2. INTRODUCTION

2.1 Background

Congress has identified resource preservation as the National Park Service's (NPS's) primary responsibility. Among the resources NPS seeks to preserve are soundscapes in which visitors have the opportunity to experience solitude or to experience nature in a state unaltered by the effects of civilization. Increased numbers of low-flying aircraft over various units of the National Park System have diminished the opportunities for solitude and for experiencing uninterrupted sounds of nature. Consequently, in 1987, Congress passed Public Law 100-91, the National Parks Overflights Act, which directed the Secretary of the Interior to conduct studies to provide information regarding the effects on resources and values of aircraft overflights of National Park units. One of the requirements of the law was that a plan be developed that would substantially restore natural quiet in the Grand Canyon. In the July 1995 Report to Congress, NPS defined "substantial restoration" in the Grand Canyon to mean "...that 50% or more of the park achieve "natural quiet" (i.e., no aircraft audible) for 75 - 100 percent of the day."[23]

Congress recognized in PL 100-91 that the need for a plan to restore natural quiet required the involvement of the Secretary of Transportation, through the Federal Aviation Administration (FAA). Working in cooperation with the NPS, the FAA designed special use airspace (SFAR 50-2) to help channel air tour routes away from sensitive areas and restore natural quiet. However, in the July Report to Congress, through use of both sound monitoring and computer modeling, the park service concluded that implementation of SFAR 50-2 had not brought a substantial restoration of natural quiet to the Grand Canyon.[24] Because the goal of substantial restoration of natural quiet was not achieved, NPS and the FAA are currently working on revising the Grand Canyon airspace so that this goal will be met in the foreseeable future.

As is typical of airspace / noise related planning efforts, computer models are the primary tool for analysis of changes to the airspace. Both NPS and FAA have used their own specially designed computer models for analyzing the noise exposure produced by changes in airspace use.[25] These models (NPS's NODSS and two versions of FAA's INM) have never been compared with sound levels produced by air tour operations over National Park settings, over a range of aircraft operating conditions. Consequently, the decision was made jointly by NPS and FAA to conduct a field measurement-based validation of the models, using third party experts (the Technical Review Committee, TRC) to provide comments and suggestions on both the methods of testing and on the results. Additionally, a fourth model, NOISEMAP Simulation Model (NMSIM) developed by Wyle Laboratories, the US Air Force and NASA, are included in the validation study. This document describes the testing and analysis of the models and the results of the analysis.

[23] U.S. Department of the Interior / National Park Service, "Report on Effects of Aircraft Overflights on the National Park System" Report to Congress, July 1995, p 182. This definition means that natural quiet is substantially restored if areas where aircraft are audible for more than 25% of the day comprise less than half the park.

[24] Ibid, p. 195.

[25] The FAA model is the Integrated Noise Model (INM), modified for use in modeling the Grand Canyon; two versions of this model are examined in this study – Version 5.1, based on A-weighted information, and the Research Version, which uses frequency-based information about each aircraft. The NPS model is the National Park Service Overflight Decision Support System (NODSS), designed and programmed specifically for park applications where audibility, significant changes in terrain elevation and shielding due to terrain features must be addressed.

2.2 Study Goals

The primary goal of this study is to:

Determine the degrees of accuracy and precision that existing computer models provide, in comparison with field measurements, in the calculation of the percent of time tour aircraft are audible in the Canyon, and calibrate one or more of these models to provide a tool for computation of air tour audibility in the Canyon[26].

The first part of the goal, determining accuracy and precision, will be termed "validation." This effort shows how well the models, when used with a common set of input variables, produce results that agree with measurements. The second part, "calibration" uses whatever techniques are deemed reasonable and scientifically supportable, to improve the agreement between measured and computed results.

2.2.1 Application First in Grand Canyon

Though the long-term goal is to have a model (or models) that predicts tour aircraft audibility (and hourly equivalent sound level) in any National Park, the present effort focuses first on model validation and calibration for the Grand Canyon. As discussed in the Background above, the need exists to design the use of Grand Canyon airspace so that natural quiet will be substantially restored. Hence, the method is applied to the Grand Canyon. Through this effort, it is expected that improved understanding of the models and of validation techniques will lead to concepts and methods for extending the use of the model(s) to other park situations.

2.2.2 Study Priorities

In-depth consideration of the goal and of several important factors related to the goal has emphasized the complexity of this model validation task. Because audibility of aircraft is the primary concern, the factors of long distance sound propagation and non-aircraft background sound levels play a significant, if not the most significant role in determining when and where aircraft are audible. Consequently, this study was designed to focus first on acquiring audibility data at different locations in the Canyon, using trained observers and specific listening / logging techniques. This study was designed to provide the most direct and efficient path to determining the accuracy and precision of the models in relation to aircraft audibility. However, it was also designed to demonstrate as efficiently as possible how well the model sound level results compare with the measured sound levels (hourly L_{eq}).

Though this study is designed to efficiently achieve the primary goal, it also provides some of the detailed types of information needed for model diagnostics. This information is provided in two ways. First, Section 9.2 presents analyses of the discrepancies (differences) between computed and measured aircraft audibilities and between computed and measured aircraft equivalent sound levels. It shows which modeled factors correlate significantly with the discrepancies between computed and measured results, and suggests which aspects of the model(s) need further investigation / improvement. Second, in Section 9.3, the study provides an analysis of only the measurement results and identifies which physical factors (such as aircraft type, wind direction, terrain, etc.) have a statistically significant correlation with the measured audibility. Such information suggests what factors the models should include if they are to produce results that agree with the measurements

[26] In addition to examining the "percent of time audible", the tour aircraft "hourly equivalent sound level" L_{eq} was also examined.

2.2.3 Failsafe Method

Because of the significant resources required to conduct this study, there was an interest in insuring that, if none of the models proved sufficiently accurate, an empirical method could be derived to estimate tour aircraft audibility in the Canyon. Though this goal is of secondary importance, and certainly of limited value as far as other parks are concerned, an empirical relationship between tour audibility and the measured variables is provided in Section 9.3. This relationship is a necessary outcome of the analysis of the measured data. In order to understand the effects of variables such as aircraft type and wind direction on the measured audibility, a statistical analysis of measured data was necessary and is reported in Section 9.3. That analysis is based on developing a mathematical relationship between percent of time audible and the various measured variables, such as number of aircraft per hour, distance from the flight corridor, etc. Even if the empirical relationship is not used, the associated analysis provides useful information for model diagnostics and improvement, should those be pursued.

The following two sub-sections first describe the primary factors considered in developing the proposed study, then outline how these factors are accounted for in the study. Section 3 provides an overview of the computer models examined, while Section 4 summarizes the overall study. Sections 5 through 8 describe the study and its results in detail. Section 9 analyzes possible factors that cause differences between measured and modeled results, while Sections 10 and 11 provide overall conclusions and recommendations for model use, model improvement and further useful analyses. Section 2.5 summarizes the chronology of the entire process.

2.3 Factors Considered in Developing the Study

2.3.1 Noise Metric

For the Grand Canyon, substantial restoration of natural quiet is the issue, and the metric of primary interest is audibility of aircraft. Each of the four models to be tested can provide a calculation of the time, or percent of time, that aircraft will be audible at locations throughout the park. Measurement of this metric requires attended monitoring; measurement, in fact, requires only an observer with normal hearing, and no equipment other than a watch and some sort of logging device such as a clipboard and pencil.

Because trained acoustics staff were present at most of the sites during measurements, tape recordings were also made simultaneously with the audibility logging conducted at these sites. These tape recordings were used to determine ambient sound levels during the measurements and equivalent sound levels, L_{eq}, of the tour aircraft, as well as to provide additional data for understanding and documenting the soundscape of the Grand Canyon.

2.3.2 Region of 25% Aircraft Audibility

Natural quiet is substantially restored when no aircraft are audible in 50% or more of the park for 75% to 100% of the day. This definition requires that the models should predict the area of the park where aircraft are audible for more than 25% of the time, and compute this area. If the area is less than half the park, then natural quiet is substantially restored. Hence, the desirable model validation outcome is to be certain to test the models in regions of the park where aircraft are likely to be audible 25% of the time.

Two approaches were taken to estimate where these areas lie, relative to the flight tracks flown. First, both INM version 5.1 and NODSS have been used previously to model Grand Canyon airspace

operations.[27] Each model computed tour aircraft audibility of 25% or more at approximately 8 to 10 miles from busy flight corridors.

Second, a simple analysis of aircraft time audible was conducted by assuming given distances from an observer to the threshold of audibility and a given aircraft speed. For example, if the threshold of audibility is 10 miles from an observer, and an aircraft flies straight through this 10-mile radius circle about the observer at 100 kts, at a distance of one mile from the observer, then the aircraft would be audible for about 10 minutes. For this situation, only one and one-half aircraft per hour could fly past the observer to be audible for 25% of the hour (15 minutes). This analysis suggests that if a corridor carries more than a few aircraft per hour (2 to 4 aircraft per hour), then the distance from the flight corridor where aircraft are audible no more than 25% of the time is quite close to the distance where aircraft are just audible. Aircraft have been noted as audible at more than 5 miles from flight corridors, and as far as 10 to 15 miles from the corridors.

Both these analyses suggest that the areas of most interest for model validation lie between 5 and 15 miles from the flight corridors. Hence, site selection focused on these areas, with measurement sites located predominantly between 5 and 15 miles from the corridor. It should be noted, however, that in order to test the models' range of capabilities, data collection sites were selected both closer to flight corridors and further from them than the anticipated location of 25% of the time audible. To ensure coverage, a few sites were chosen to be beyond the distances of tour aircraft audibility.

2.3.3 Aircraft Altitudes

Tour aircraft flying over the Grand Canyon are at altitudes of no more than approximately 7000 feet above the Canyon floor, and typically no more than 1000 to 2000 feet above the Canyon rim. For an observer 10 miles distant from the flight corridor, these altitudes place the aircraft at roughly 2 to 8 degrees above the horizon, relative to the observer. Thus, it was expected that propagation effects caused by wind and / or temperature gradients might strongly influence measured results so that during measurements meteorological data were collected from both temporary and permanent "met" stations.

2.3.4 Aircraft Audibility and Ambient Sound Levels

"Audibility" as used in this study begins at the instant that an attentive human listener can detect the presence of the sound produced by a tour aircraft and lasts as long as the listener continues to hear the aircraft. Though the audibility of a source can vary from listener to listener, on average, humans without significant hearing loss are able to identify the presence of a source in a given background sound environment at similar sound levels. Whether or not a tour is audible is determined by both the sound level of the tour aircraft and by the sound level of the ambient or non-tour aircraft sound levels. These concepts, the mathematical form used to compute audibility, and the measured performance of the field staff that collected the audibility data are presented in detail in APPENDIX C, page 167.

The two primary factors that determine whether or not an aircraft is audible at a given location are the aircraft's sound level and the sound level of the surrounding non-aircraft background sounds (referred to here as the ambient sound levels). Any model used to compute aircraft audibility must incorporate both these variables. Though aircraft noise models traditionally include some type of aircraft sound level database, they usually do not provide for ambient sound levels. The four models

[27] The INM was used in the Final Environmental Assessment of the Grand Canyon Airspace, footnote 9, and NODSS was used to provide data for the NPS Report to Congress.

to be validated here can all incorporate ambient sound level information in their calculation of aircraft audibility.

Hence, each model needs to have ambient sound levels identified for all locations where computations of aircraft audibility are required. Obviously, ambient levels are variable over time and from location to location within a park; rigorous full quantification in both time and space would be extremely difficult and is judged here as impractical. The NPS and FAA have been developing methods based on sampling to quantify ambient levels, and continue to refine these levels for the Grand Canyon.[28] This study uses two types of ambient levels: first, the "measured ambient" sound levels, so called because they are measured (using tape recordings) simultaneously with collection of audibility data; 2) the "EA ambient" used in the Environmental Assessment of the Grand Canyon air space changes.[29] Throughout this report, the analyses separately identify and discuss the modeled audibility results as produced using either the "measured ambient" or the "EA ambient".

2.4 Study Design

Consideration of these four factors as well as discussions among NPS and FAA staff, the TRC members, and consultants resulted in the study reported here. In general, this study consisted of acquiring sufficient data during four days in the Canyon to permit a statistically significant comparison of the tour aircraft audibility and sound level data that were measured with computer model estimates of these audibilities and sound levels.

The study included data collection at 39 different audibility sites, five temporary and two permanent meteorological sites, and one aircraft source level site in the Grand Canyon, with about 300 hours of audibility data, supplemented by about 200 hours of tape recordings. During data collection, ten different teams conducted measurements: eight dedicated to collecting tour audibility data and tape recordings in the Canyon; one to measuring tour aircraft source sound levels near the tour route, and one to collecting meteorological data. National Park Service staff provided all logistics support for transportation of instrumentation, camping gear, food and water into the measurement sites.

Data reduction was also split among different groups. Staff from Volpe and HMMH reduced the aircraft source sound levels; HMMH collected and distributed the model input information and reduced the audibility data. Volpe ran the INM versions, NPS ran NODSS, and Wyle ran NMSIM. HMMH provided the statistical analysis and most of the documentation.

2.5 Study Process

The study process has taken over two years of effort by all participants. Table 5 provides a summary chronology of the process. Notes and appendices provide additional information, as indicated.

[28] See "Natural Ambient Sound Levels for Use in Noise Modeling of Grand Canyon," Memo from N. Miller to NPS, Dec. 2, 1998, and "Addendum: Natural Ambient Sound Levels for use in Noise Modeling of Grand Canyon NP," Memo from N. Miller to NPS, Feb. 5, 1999.

[29] See footnote9.

Table 5. Chronology of Model Validation Study

Date	Description of Accomplishment
October 1998	Development of Model Validation Study Plan Authorized; Formation of Technical Review Committee Authorized.
January 1999	First Draft of Plan Provided.
January – July 1999	Technical Review Committee Formed. [A]
April 1999	Second Draft of Plan Provided.
July 1999	Third Draft of Plan Provided.
August 15 – 18, 1999	First Meeting of Team (NPS, FAA, Volpe, Wyle, HMMH) with TRC to Review Study Plan (at Grand Canyon); [B] Public Presentation of Study Plan. [C]
August 31, 1999	Final Study Plan Submitted to NPS.
September 7-15, 1999	Data Collection in Canyon.
October 1999 – April 2000	Assemble All Data.
April 2000	Data Reduction and Analysis Authorized.
April – September 2000	Determine Source Levels; Provide Model Input to All Modelers.
September 2000 – March 2001	First Run of Models; NMSIM using “soft ground” assumption; INM models used without compression algorithm. Initial Statistical Analysis of Computed v. Measured Audibilities.
March 28, 2001	Second Meeting of Team with TRC to Review Draft Results. [D]
May 2001	Additional Analysis, Documentation of Study to Date Authorized.
May – July 2001	NMSIM rerun using “hard ground” assumption as recommended by TRC INM rerun using compression algorithm.
September 2001	First draft of report provided; Internal NPS, FAA review of report.
November 2001	NMSIM rerun using corrected time delays for hourly data (Section 6.1.1).
February - March 2002	Second draft of report provided; TRC, NPS, FAA review of Documentation
June 2002	Third draft of report provided; TRC, NPS, FAA review of report
November 2002	Final draft of report provided; NPS final review.

Notes to Table 5:

[A] See APPENDIX A for membership and Charter.

[B] See Appendix B.1 for agenda and attendees.

[C] See Appendix B.2 for attendees.

[D] See Appendix B.3 for agenda and attendees.

3. OVERVIEW OF FOUR MODELS TO BE VALIDATED

3.1 Introduction

This section provides brief descriptions of the four models examined - their basic concepts, their inputs and outputs.

3.2 Computer Models Tested

The initial intent of this study was to test one FAA and one NPS model, each of which has been used to analyze the audibility of tour flights over the Grand Canyon. However, in order to include a broader range of types of aircraft noise computer models, a third model has been added to this plan, and FAA has included a second version of its program. The current FAA model is derived directly from the Integrated Noise Model (INM, Version 5.1), the primary aircraft noise model used in the U.S. for analysis of civil aviation. The NPS model, the NPS Overflight Decision Support System (NODSS) was designed specifically for use in calculating the audibility of aircraft in a National Park setting. A third model, NOISEMAP Simulation Model (NMSIM), developed by Wyle Laboratories, the US Air Force and NASA, is based on NOISEMAP, the US Department of Defense model used for analysis of military aircraft operations. Additionally, a second version of the INM, the Research Version, has been developed by the FAA and Volpe and is included in the testing. This section discusses the basic approach used by each model.

The approach is to use each model as it is intended to be used, with no substantive changes to its basic computations. Some effort was necessary to update or improve basic input databases (such as revising the aircraft noise database of the INM, or the aircraft spectral data used by NODSS and NMSIM) or to modify presentation of outputs (as the INM output was modified to provide contours of "time above" a threshold, and as NMSIM was modified to compute audibility). The goal, as stated, is to examine current models, rather than to enter into an open-ended model design process. The Research Version of the INM, however, represents the beginning of the process to incorporate detailed spectral information into the INM calculations of audibility.

3.3 General Description of the Models

3.3.1 INM

3.3.1.1 Version 5.1

The INM is the FAA developed, internationally used aircraft noise computation model that runs on an IBM PC or compatible.[30] It does calculations based on A-weighted aircraft sound levels, adding up sound energy from the different segments of the aircraft flight tracks. It has gone through many versions, each an improvement in accuracy or ease of use. It is, in the United States, the model of choice for analysis of civil aviation noise in the vicinity of airports. For computation of sound energy based metrics (equivalent sound level, L_{eq}), version 5.1 of the INM uses an aircraft-specific database of A-weighted information: SEL (Sound Exposure Level) *versus* slant distance to the aircraft. The SEL for any aircraft depends upon thrust / power, and the INM sums the sound energy from all flight operations at a grid of points on the terrain surface. The grid of points is used to construct contours of equal equivalent levels. To compute "time above" a given sound level threshold for a given aircraft flight, the INM assumes a dipole directivity pattern for the aircraft, and

[30] Olmstead, *et al*, "Integrated Noise Model (INM) Version 5.1 User's Guide," FAA-AEE-96-02, December 1996.

constant speed, and uses the difference between the aircraft produced SEL and Lmax to compute the time the level at the receiver is above a user specified level.[31] In this study, to compute the time aircraft are audible, the user specified level is chosen to approximate the threshold of audibility, but in terms of A-weighted levels. In its calculations, the INM accounts for differences in terrain elevations relative to aircraft flight tracks, but does not include shielding effects of terrain.

3.3.1.2 Research Version

The Research Version of the INM utilizes the standard INM database and aircraft-specific spectral data measured at the Source Site to calculate noise metrics. As with Version 5.1, this version also runs on an IBM PC or compatible. Aircraft flight paths are evaluated on a segment-by-segment basis to calculate both sound level and audibility metrics. For the calculation of audibility, the spectrum at the time of maximum sound level is A-weighted, corrected back to the source at a reference distance, and then corrected for both spherical spreading and atmospheric absorption to the appropriate slant distance for the given segment. The final, corrected spectrum is then evaluated using traditional detectability calculations, using a 10 log (d') value of 7 dB for audibility.[32] The audibility time is determined for each flight segment, then all times summed to give the total audibility for a given aircraft flight/group of flights.

3.3.2 NODSS

NODSS is an omni-directional point source model, does frequency dependent calculations of audibility, and accounts for terrain elevations and shielding effects.[33] It runs on a Sun Ultra 1 Unix workstation. It steps the aircraft along a user-defined track in increments (nominally 300 m, but the stepping distance can vary, depending on the geometry involved, to strike a balance between accuracy and computation time). The model starts the airplane at the point of closest approach (PCA), and works along the flight path in one direction until the sound level drops sufficiently so as not to be of further interest for the parameters being calculated (but making sure that the sound level is not dropping just because the airplane is momentarily hiding behind a terrain barrier). NODSS then places the aircraft back at the PCA and steps it in the other direction. Thus a complete time history of 1/3 OB spectra is calculated for the full overflight, and from this time history, the various metrics are computed. NODSS uses the full detectability calculations of 10 log (d') equal to 7 dB to determine audibility.

3.3.3 NMSIM

NMSIM is a simulation model that computes aircraft sound level time histories as experienced on the ground.[34] It does frequency dependent calculations and accounts for aircraft directivity, and terrain elevations and shielding effects. It "flies" the aircraft through a user-specified flight path, and computes the noise at user-specified points on the ground. The aircraft source noise is based on

[31] Olmstead, J.R., *et al*, "Integrated Noise Model (INM) Version 6.0 Technical Manual", Federal Aviation Administration, FAA-AEE-02-01, January 2002, Appendix C.

[32] 10 Log (d') is a measure of a signal's relationship to the background "noise. For a complete description of audibility, detectability and the associated mathematics, see APPENDIX C, page167.

[33] Reddingius, N.H., "User's Manual for the National Park Service Overflight Decision Support System," BBN Report 7984, 10 May 1994.

[34] Ikelheimer, B., *et al*, "Noise Model Simulation (NMSIM) Beta Test Version," Wyle Report WR 01-16, May 2002. Note that this report documents the configuration and use of the NMSIM version that is currently being converted to the Windows environment. NMSIM Version 2.3A used for this study is the original DOS version of the program. This manual, however, is, useful for a general description of how NMSIM functions.

original "NOISEFILE" (from NOISEMAP) data, and provides sound level, 1/3-octave spectra, and directivity information. Propagation from the aircraft to the ground is performed on a path-by-path basis. Terrain effects are included, using the algorithms that are employed in NOISEMAP 7.0. When NMSIM is run, it computes complete time histories of 1/3-octave spectra at each receiver point. Any noise metric can then be computed from these time histories. To determine audibility, NMSIM uses the full calculation of 10 log (d') equal to 7 dB. NMSIM runs on an IBM PC or compatible. It has an interactive mode, in which the user operates it from a map display showing the terrain, flight path and receivers. The noise from any point on the flight path can be examined, as well as noise from the complete flight. There is also a batch mode, where NMSIM can generate any of the noise quantities it computes at points located on a defined rectilinear grid.

Figure 18, Figure 19, and Figure 20 present schematic comparisons of the models. In these figures, the input variables are shown on the left, the computational modules in the middle, and the output variables on the right. Schematically the INM Research Version is similar to INM Version 5.1 except that all sound level information is in ⅓ octave bands.

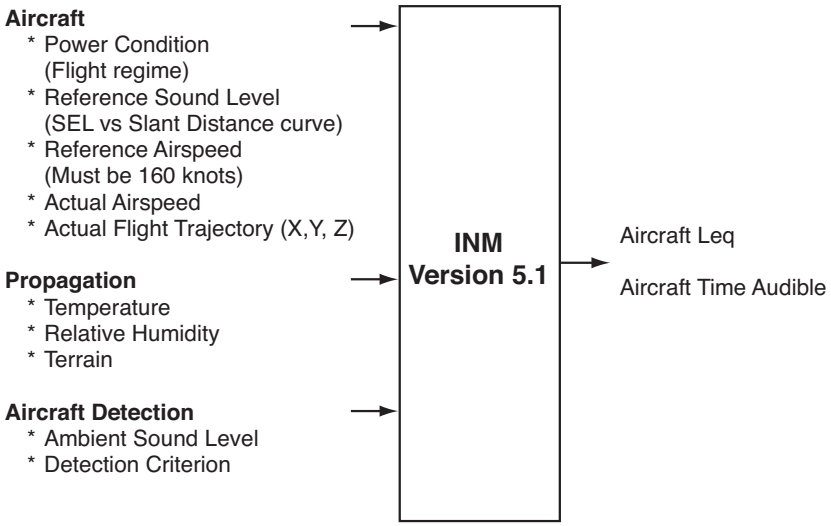


Figure 18. Schematic of INM Version 5.1

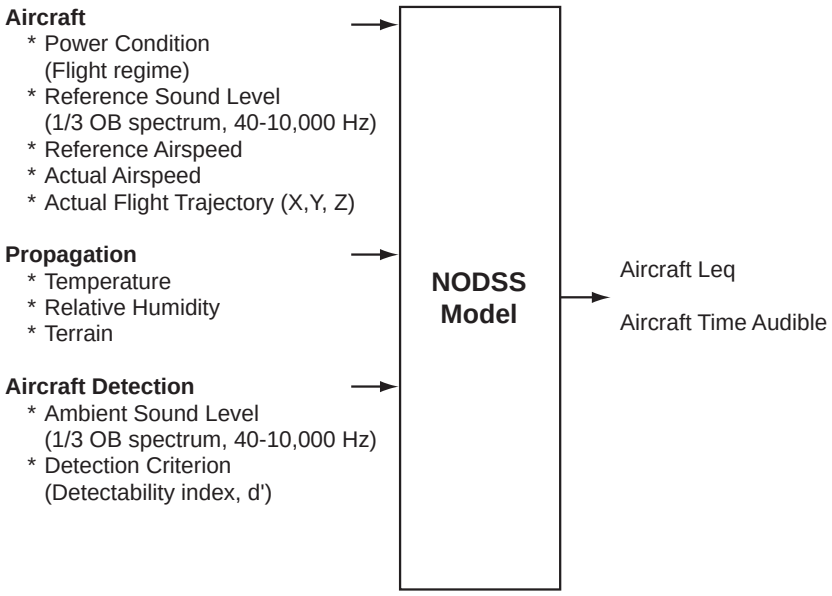


Figure 19. Schematic of NODSS

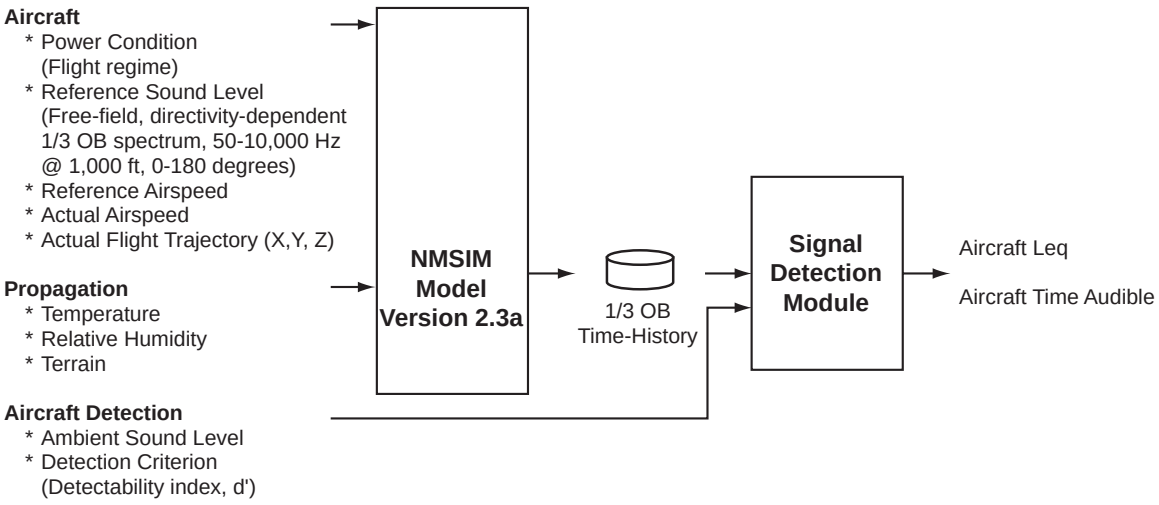


Figure 20. Schematic of NMSIM

3.3.4 Note on Computation of Percent of Time Aircraft are Audible

Each model computed the percent of each hour that tour aircraft are audible. The different models used slightly different approaches. The INM versions and NODSS computed percents by first calculating the seconds of audibility per aircraft and summing these times together and dividing the sum by seconds in an hour. Since this approach does not account for the possibility of overlapping audibilities when aircraft are close together, the total percent time audible for an hour sometimes exceeds 100%. To convert these results to realistic percents, a "compression" algorithm was used. This algorithm is based on measurements made in various Canyon locations in 1992, and is given in APPENDIX J, page 243.

3.4 Model Inputs

Each model's input requirements are described below. Some of these requirements affected data collection and reduction procedures, while others are either built into the models, or are available from public databases (such as terrain information[35]).

Each model requires three basic types of input information: 1) information about the aircraft position and noise produced; 2) information related to sound propagation; and 3) information related to when an aircraft will be audible. There is considerable overlap among the model input requirements, but some differences as well. The main difference is between the INM Version 5.1 and all other models. Only INM 5.1 makes all computations using A-weighted sound levels. Detection theory (i.e., when a sound becomes audible, see APPENDIX C, page 167) is based on frequency information. "A-weighting" is a method for combining the sound levels in all frequencies,

[35] NODSS - "Digital Line Graphs from 1:100,000-Scale Maps, Data Users Guide 2," United States Department of the Interior U.S. Geological Survey, Reston, VA, 1989.
INM - Gulding, *et al*, "Integrated Noise Model (INM) Version 6.0 User's Guide", Washington, D.C, Federal Aviation Administration, FAA-AEE-99-03, September 1999, Appendix A.
NMSIM - USGS 30 meter DEM's , mosaicked together to cover the Special Flight Rules Area.

much as human hearing does, to produce a single number. Hence, for use of INM 5.1, an estimate was needed of the A-weighted level at which a tour aircraft becomes audible. Previous analysis has shown that, for a selection of Grand Canyon ambients, tour type aircraft (helicopters and propeller aircraft) on average can become audible when the A-weighted aircraft sound level is 8 dB below the ambient sound level.[36] For the other models (including the Research Version of the INM), which use spectral information, the threshold of audibility of tour aircraft was assumed to occur when the detectability level, referred to as d-prime (D') or 10 log (d'), has a value of 7 dB. (See Appendix C.2.2, page 167.)

3.4.1 INM

3.4.1.1 Version 5.1

Table 6 summarizes the input variables for INM Version 5.1. The left-hand column in the table describes the variable and the right-hand column identifies its purpose. In standard (airport related) computations, a "threshold" may be used, rather than the "ambient". The INM will calculate the amount of time a threshold (in dBA) is exceeded. In this application to the Grand Canyon, the threshold used is the level at which aircraft are assumed to become audible. For this use, the threshold is determined by first identifying the ambient level, then adjusting it by the detection criterion. So, for example, if the ambient is 25 dBA, and the detection criterion is ambient minus 8 dB, then the INM threshold is set to 17 dBA. Each of these variables is a single-valued number, in decibels.

3.4.1.2 Research Version

Table 7 summarizes the input variables for the Research Version of the INM model. The left-hand column in the table describes the variable and the right-hand column identifies its purpose. These variables are the same as those required by version 5.1 of the INM (though different terminology is used here), with the exception of the spectral data. The Research Version uses actual spectral data measured at the Source Site, whereas the current public release version of the model, version 6.0c uses "Spectral Classes" which are generalized spectral forms for different aircraft types. The detection criterion for the Research Version is 10 log (d') = 7 dB.

3.4.2 NODSS

Table 8 summarizes the input variables for the NODSS model. All of the Aircraft variables are input to the model via an ASCII text file. For tour audibility calculations, the ambient level is determined from a pick list in which several ambient 1/3-octave band spectrum shapes are available. A shape is picked and the A-weighted sound level of that spectrum is specified in order to calculate the individual 1/3-octave band sound levels from that shape. The choice of shape depends upon the vegetation zone applied to each location in the Canyon for which tour audibility is calculated (see APPENDIX G page 213.

The atmospheric conditions considered by the model for air absorption calculations are presently hard-wired for standard day conditions (59F, 70 % relative humidity, and sea level atmospheric pressure). The detection criterion is also hard wired for a detectability level (D') of 10 log (d') = 17 dB. By reducing the ambient sound levels modeled in NODSS by 10 dB, NODSS computed results behaved as if the detection criteria were that used by the other models of 10 log (d') = 7 dB.

[36] See "Review of Scientific Basis for Change in Noise Impact Assessment Method Used at Grand Canyon National Park," National Park Service, January 2000.

NODSS input was altered for computation of tour aircraft equivalent sound levels. NODSS normally calculates equivalent sound level as a total level that includes both aircraft sound and ambient sound. Since only aircraft equivalent sound levels were needed, for these computations the ambient was set equal to –80 dB so that ambient sound would not contribute to the computed equivalent level.

Table 6. Input Variables for INM Version 5.1

VARIABLE	PURPOSE
Aircraft	
Name	Documentation
Power Condition (flight regime)	Computation
Airspeed	Computation
Number of aircraft per time	Computation
Sound Exposure Levels (SEL) *versus* slant distance	Computation
Flight path (X, Y, Z)	Computation
Ambient	
Sound Level (dBA)	With Detection Criterion, determines threshold used.
Terrain	
Commercial terrain database	Computation
Atmospheric	
Temperature	Computation
Relative humidity	Computation
Detection Criterion	
Difference between aircraft and background sound levels (dB) for threshold of audibility or noticeability. Currently use 8 dB below ambient (audibility)	With Ambient, determines threshold used.

Table 7. Input Variables for Research Version of INM

VARIABLE	PURPOSE
Aircraft	
Name	Documentation
Power Condition (flight regime)	Computation
Airspeed	Computation
Number of aircraft per time	Computation
Aircraft-specific spectral data measured at Source Site	Computation
Flight path (X, Y, Z)	Computation
Ambient	
1/3 Octave band sound levels (50 – 10,000 Hz)	With Detection Criterion, determines threshold used.
Terrain	
Terrain database	Computation
Atmospheric	
Temperature	Computation
Relative humidity	Computation
Detection Criterion	
d' signal detection value for threshold of audibility. Used 10 log (d') = 7 dB	Computation

Table 8. Input Variables for NODSS

VARIABLE	PURPOSE
Aircraft	
Name (manufacturer's designation)	Documentation
Name (common)	Documentation
Power Condition (flight regime)	Computation
Airspeed	Computation
Number of aircraft per time	Computation
Propeller Speed	Documentation
1/3 Octave band sound levels (40 – 10,000 Hz) @ standard day conditions and reference distance of 1,000 ft.	Computation
Flight path (X, Y, Z)	Computation
Ambient	
1/3 Octave band sound levels (40 – 10,000 Hz) plus A-level of spectrum*	Computation
Terrain	
Terrain database	Computation
Atmospheric	
Temperature (default = 59F)*	Computation
Relative humidity (default = 70%)*	Computation
Detection Criterion	
d' signal detection value for threshold of audibility. Current default 10 log (d') = 17dB (input adjusted to yield 7dB, see text)	Computation
*Software modification required to allow user-specified values.	

3.4.3 NMSIM

Of all four models, NMSIM requires the largest amount of aircraft sound level information (see Table 9). The aircraft sound levels must be described as a series of a directivity-dependent 1/3 octave band sound level spectra under free-field conditions (i.e., absent any ground reflections) at a reference distance of 1,000 feet. These spectra were derived from tape recordings made during measurements. The model applies algorithms that account for atmospheric absorption, terrain barriers, and ground impedance (reflections) in addition to inverse-square spherical spreading in calculating a 1/3 octave band spectrum at the receiver location. The sound level directivity pattern and the aircraft ground speed are used to generate at the receiver's location a 1/3 octave band time history, the maximum A-weighted sound level, and the Sound Exposure Level (SEL) of the overflight.

Table 9. Input Variables for NMSIM

VARIABLE	PURPOSE
Aircraft	
Name	Documentation
Power Condition (flight regime)	Computation
Airspeed	Computation
Time of each aircraft flight	Computation
Free-field 1/3 Octave band sound levels (50 – 10,000 Hz) at distances of 1,000 ft. (with air absorption removed), for a range of in-flight directivity angles of 0^o (nose) to 180^o (tail). Axial symmetry of the noise source is assumed.	Computation
Flight path (X, Y, Z)	Computation
Ambient	
1/3 Octave band sound levels (50 – 10,000 Hz)	Computation
Terrain	
Terrain database	Computation
Atmospheric	
Temperature	Computation
Relative humidity	Computation
Detection Criterion	
d' signal detection value for threshold of audibility. Used 10 log (d') = 7	Computation

3.5 Model Outputs

Model outputs for the purpose of this study are virtually identical for each of the four models. They include the length of time tour aircraft are audible, (converted for this study to percent of time audible), and tour aircraft hourly equivalent sound level, L_{eq}. As mentioned in Section 3.3.4, the INM versions and NODSS do not account for overlapping of aircraft audibility times, and the results of these models were "compressed" using the equation of APPENDIX J, page 243.

4. STUDY APPROACH OVERVIEW

The study approach is based on the concept that aircraft noise model validation is best accomplished by comparing computed results with measured results. The model computations are based on the actual operations that were measured, and are done for the exact locations where the measured results were acquired. This section provides a brief overview of each step of this study, and the following sections, together with associated appendices provide detailed descriptions and summaries of the data and of the results.

4.1 Data Acquisition

Data acquisition was accomplished with ten separate teams: 1) eight four-person teams collecting audibility data at 39 sites and acoustic data (tape recordings) at 19 of those sites to the east and west of the Zuni Point corridor; 2) one four-person team near Papago Point on the south rim, under the Zuni Point corridor, measuring tour aircraft source sound levels and speeds and logging tour times and aircraft types; 3) one team overseeing the collection of meteorological data at five temporary sites. (Details in Section 5.)

4.2 Data Reduction

Data collected in the Canyon were reduced to forms useful for two purposes: 1) for input to the three computer models; 2) for analysis of measured *versus* computed tour aircraft audibility and tour aircraft equivalent level. The primary data needed for modeling included number and type of tour aircraft per hour, their speeds and sound levels; the average temperature and relative humidity during the measurements; the specific coordinates of all measurement locations; and ambient sound levels at each site. Data used for analysis included percent of each hour tour aircraft were audible at each site; aircraft hourly equivalent levels, L_{eq}, for each hour for each site where tape recordings were made; site parameters such as distance to the flight corridor, angle of corridor visible, site elevation; meteorological data from the "met" station nearest each site, including average wind direction and speed, relative humidity, temperature and atmospheric pressure during each hour. (Details in Section 6.)

4.3 Modeling

Using the input data, each of the four models was run to compute tour aircraft hourly percent time audible and hourly equivalent level for each measurement site during the measurements. As discussed below in detail in APPENDIX C, page 167, computation of audibility requires input information about both the sound level of the aircraft, and about the sound level of the ambient sounds. Each site has been modeled and analyzed using three different ambients: 1) the ambient used in the Environmental Assessment of the changes in tour routes[37] termed here the "EA ambient"; 2) the ambients measured at each site, the "measured ambient"; 3) and the measured ambient plus 10 dB. The last was used to qualitatively assess the sensitivity of the computations of each model to the selection of the ambient level and is presented in APPENDIX H, page 233 in graphical form only. The models also computed the tour aircraft hourly equivalent sound levels, L_{eq}, for each site, for each hour. Since the computed aircraft L_{eq} are independent of the ambient sound level, each model computed one aircraft L_{eq} for each site for each hour of measurement. (Details in Section 7.)

[37] "Special Flight Rules in the Vicinity of Grand Canyon National Park, Final Supplemental Environmental Assessment" Federal Aviation Administration, February 2000.

4.4 Data Analysis

Three types of analyses were performed on the measured and computed results.

First, the computed results were compared with the measured results for both audibility and for tour aircraft L_{eq}, (Section 8). The comparisons determine the overall error, the accuracy, precision and contour error of the models.

Second, the "discrepancies" or differences between the computed and measured results are analyzed for both the audibilities and for the equivalent levels (Section 9.2). These analyses identify which physical factors are most statistically significant in being correlated with the differences between computed and measured results. Model improvements, if any are desired, should first focus on the factors identified by this analysis. Section 11.2 provides recommendations for model improvements.

Third, the measured results were analyzed to identify which physical factors correlated with the measured audibilities. The goal was to provide insight into how important factors such as aircraft type, wind, and temperature were in relation to the measured results. Such an analysis primarily provides useful information to model development / refinement by identifying which factors the models should incorporate. Additionally, this analysis yields a strictly empirical relationship between audibility of tour aircraft in the Canyon and the various physical parameters. (Details are provided in Section 9.3, page 138.)

5. DATA ACQUISITION

Data acquisition focused on measuring specific variables at specific sites. This section describes the primary considerations of site selection, lists the sites and instrumentation used, and describes the data acquisition methods, and Table 10 provides the general schedule that was followed during the field work portion of data acquisition.

5.1 Site Selection

The goal of the study was to collect audibility data and sufficient associated information to permit the four models to compute the tour audibilities at the measurement sites for the measured conditions. In order to insure that the situations measured and modeled were as free as possible of undue complications, several considerations were used in selecting measurement sites. Air tour operations were to be measured primarily in level flight to correspond with the predominant flight condition. Measurements should also be made to either side of the flight corridor so that, from the corridor to the sites, both upwind and downwind conditions would be measured, if they occurred. Sites should be located both on the rim and in the Canyon to represent the wide range of Canyon conditions. Measurement sites should not be affected by the sounds from other air tour corridors. Some sites should be located distant enough from the corridor to be beyond the limit of tour audibility, but all sites should be accessible with less than one day of travel. Also, it was judged to be useful if air tour traffic were not too heavy, so that there would be periods when only one tour was audible and thus provide as simple a situation as possible for model diagnostics.

Table 10. Data Acquisition Schedule

Date	Description of Accomplishment
September 7, 1999	Teams travel to Canyon
September 8, 1999	Training; equipment assembly for sling loads to sites
September 9, 1999	All teams hike / drive to assigned sites, set up camp, unpack equipment
September 10, 1999	First full day of data collection
September 11, 1999	Second day – no significant data collection due to storms
September 12, 1999	Third full day of data collection
September 13, 1999	Fourth full day of data collection
September 14, 1999	Teams hike / drive out of Canyon
September 15, 1999	Teams return home.

The decision was made to measure the air tours using the Zuni Point Corridor, see Figure 21. Most tours using this corridor travel from south to north, (counterclockwise in the figure) are moderate in number, and include both helicopter and fixed wing aircraft at level flight. Being essentially unidirectional meant that keeping full account of every air tour was simplified; one observation point (called the Source Site) would be adequate to track the times and types of all flights. The moderate numbers meant that there would be periods when only one aircraft was audible, providing data that could be used for model diagnostics. The terrain permitted measurement sites to either side of the Zuni Point Corridor (to the east or west), on the rim and in the Canyon, and up to distances that

would virtually ensure being beyond the maximum distance of air tour audibility. All sites could be accessible in less than one day of travel, either by foot or in an automobile.

Figure 22 shows the general locations of the acoustic sites, the meteorological sites and the source site in relation to the tour route. (Sites 9A, 9C, 9D and 9E are too far to the east to show on this figure.) APPENDIX D, page181 presents figures that show in detail, the locations of the sites, while Table 11 gives site groups, specific coordinates, and type of instrumentation used / data collected in addition to audibility logging, see also Section 5.2 below. The sites are numbered by general geographic groupings. The final letter identifies the staff making the measurements: h=HMMH, n=NPS, v=Volpe. The third column identifies the group into which each specific site was placed for purposes of the site group analysis. Note also that Site 5A, Cape Final, together with the Source Site, collected tour aircraft fly-by times for calculation of aircraft speeds for use in modeling.

Note that Figure 22 shows the seven meteorological sites that provided data for comparison with computed and measured results. Two of these sites used are permanently installed in the Canyon – Abyss and Hance – while the remaining five were established solely for the time of this model validation data collection.

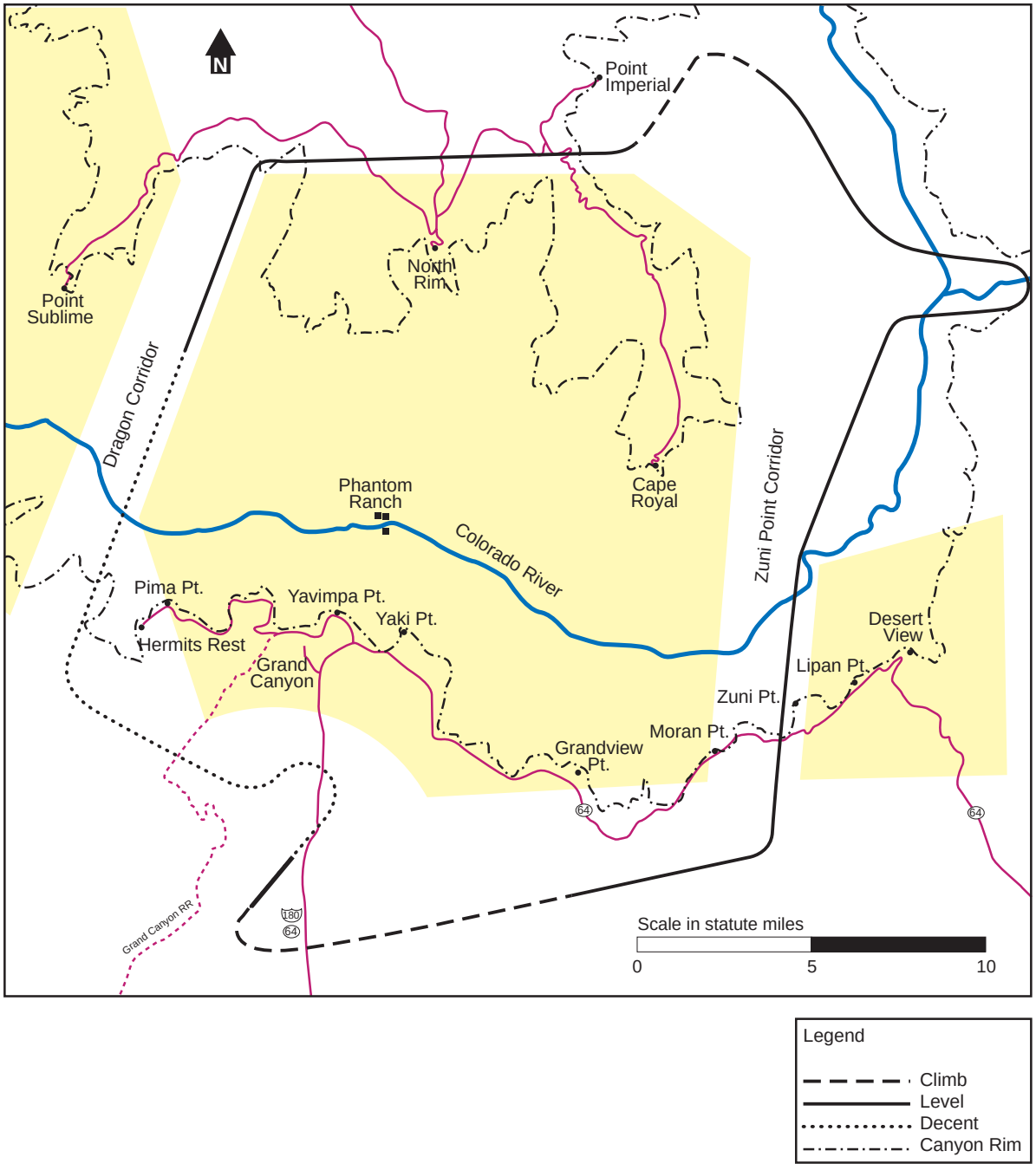


Figure 21. General Tour Route and Canyon Features

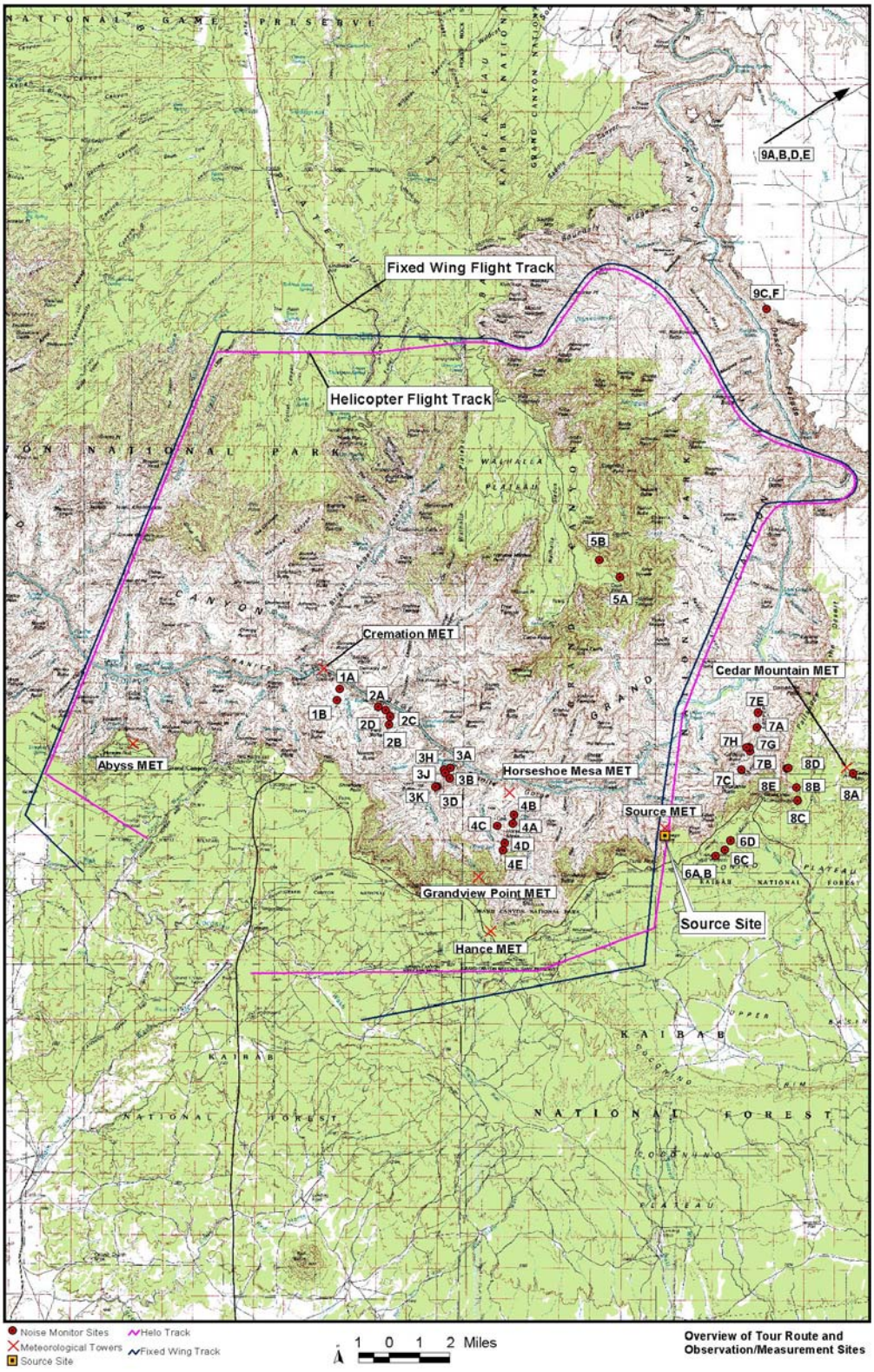


Figure 22. Data Acquisition Sites

Table 11. Audibility Sites, Locations and Additional Data Collected

Site	Name	Site Group	Lat (deg)	Long (deg)	Elev. (ft)	Digital Tape	Low Noise Mic.	Low Noise Screen
1Ah	Tipoff / Cremation	1All	36.09196	112.07306	3680	X	X	X
1Bn		1All	36.08677	112.07472	3640			
2Ah	Lone Tree	2All	36.08361	112.05055	3810	X	X	X
2Bn		2All	36.07533	112.04402	3660			
2Cn		2All	36.07909	112.04334	3750			
2Dh		2All	36.08200	112.04628	3720	X	X	X
3Av	Grapevine	3North	36.05496	112.00811	3650	X		X
3Bv		3North	36.05017	112.00846	3560	X		X
3Dv		3South	36.04639	112.01556	3580	X		X
3Hn		3North	36.05432	112.01180	4110			
3Jn		3North	36.05255	112.01115	4010			
3Kn		3South	36.04601	112.01647	3630			
4Ah	Horseshoe Mesa	4North	36.02893	111.97141	4870	X		
4Bn		4North	36.03313	111.97070	4890			
4Cn		4North	36.02794	111.98042	4900			
4Dn		4South	36.02964	111.97598	4820			
4En		4South	36.01624	111.97689	5140			
5Av	Cape Final - Rim[38]	5All	36.14634	111.91015	7960	X	X	X
5An		5All	36.14634	111.91015	7960			
5Bv	Cape Final - Interior	5All	36.15443	111.92227	8040	X	X	X
5Bn		5All	36.15443	111.92227	8040			
6Av	Desert View	6All	36.01460	111.85278	7210	X		
6Cn		6All	36.01750	111.84778	7240			
6Dn		6All	36.02181	111.84423	7290			
7Ah	Tanner Trail	7All	36.07572	111.82944	4270	X		
7Bh		7All	36.06427	111.83354	5570	X		
7Ch		7All	36.05561	111.83826	5530	X		
7En		7All	36.08279	111.82885	3970			
7Gn		7All	36.06615	111.83379	5370			
7Hn		7All	36.06606	111.83560	5620			
8Ah	Cedar Mountain	8Mtn	36.05417	111.77306	7010	X	X	X
8Bn	Bone Cache	8Ridge	36.04731	111.80622	6760			
8Cn	Switchbacks	8Ridge	36.04111	111.80547	7010			
8Dh	Palisades - Rime	8Ridge	36.05639	111.81208	6940	X	X	X
8En	Palisades – 100 yds back	8Ridge	36.05656	111.81086	6940			
9Av	Navajo – 15 mi from Zuni	9Far	36.37583	111.64067	6060	X	X	X
9Bn	Navajo – 11 mi from Zuni	9Far	36.34600	111.69283	6060			
9Cv	Navajo – 2 mi from Zuni	9Near	36.27417	111.82600	6010	X	X	X
9Dv	Navajo – 11 mi from Zuni	9Far	36.34283	111.69000	6060	X	X	X
9En	Navajo – 11 mi from Zuni	9Far	36.34383	111.69067	6060			
9Fn	Navajo – 2 mi from Zuni	9Near	36.27417	111.82600	6010			

[38] Also used to collect tour aircraft position data for use in computing average tour aircraft speeds, see Section 5.2.5. Note that 5Av and 5An are the same site, measured by different people, as are 5Bv and 5Bn.

5.2 Methods and Instrumentation

5.2.1 Audibility of Tour Aircraft

Logging of source audibility in parks (now referred to as Observer Based Source Identification Logging, or OBSIL) has been developed and conducted in parks since 1992. It was developed for NPS to identify in a controlled, orderly way, when different sounds, whether natural or human produced, are present and audible in a park setting. This logging process has been applied in at least five studies, including two for NPS, two for the FAA, and one for the U.S. Air Force.[39]

In general, for this study the observer sat quietly, some distance from any sound recording microphone, should one be present, and used a palmtop computer to track the "acoustic state" as it occurred. A software program permitted the observer to use the palmtop to enter into a spreadsheet both the exact time a source was heard, and the type of source heard. The primary goal for the observer was to identify and log onset and offset of tour aircraft sound. It is important to emphasize that this goal meant that if a tour aircraft were audible, and a different sound occurred, such as a high altitude jet overflight, the tour aircraft should continue to be the logged source as long as it was audible. This approach was used because the comparison of the measured data was to be made with computed tour aircraft audibility. If a tour were flying the corridor, the model would compute it to have some audibility. If the observer instead logged a jet, because it was also audible, then an incorrect comparison between computed and measured tour audibility would result.

Figure 23 shows the key overlay used on the palmtops, Hewlett Packard 200LX's. The observer would first press the <ALT>"Time" keys when any change of acoustic state was heard. Then, usually after briefly listening to identify the new source with certainty, the type of source was identified using <ALT> and the appropriate source key. For "Prop", "Helo", or "Pr/He", the observer would then press either <ALT>"Tour" or <ALT>"Other" depending on whether or not the aircraft were flying in the corridor.

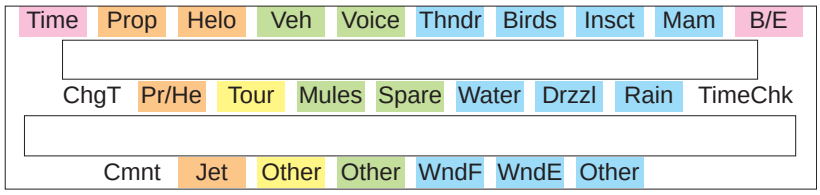


Figure 23. Key Identifiers Used for Observer Logging

[39]
- Horonjeff, *et al*, "Acoustic Data Collected at Grand Canyon, Haleakala and Hawaii Volcanoes National Parks," HMMH Report No. 290940.18, NPOA Report No. 93-4, August 1993.
- Anderson, *et al*, "Dose-Response Relationships Derived from Data Collected at Grand Canyon, Haleakala and Hawaii Volcanoes National Parks," HMMH Report No. 290940.14, NPOA Report No. 93-6, October 1993.
- Fleming, *et al*, "Development of Noise Dos/Visitor Response Relationships for the National Parks Overflight Rule: Bryce Canyon National Park," DOT-VNTSC-FAA-98-6, July 1998.
- Fleming, *et al*, "Ambient Sound Levels at Four Department of Interior Conservation Units in Support of Homestead Air Force Base Supplemental Environmental Impact Statement," FAA-AEE-99-02, June 1999.
- Miller, *et al*, "Mitigating the Effects of Military Aircraft Overflights on Recreational Users of Parks," HMMH Report No. 294470.04, AFRL-HE-TR-2000-0034, DTIC ADA379467, July 1999.

Because so many teams were conducting the monitoring, members from every team received both classroom and field training in this logging procedure (see Table 10 on page 51, September 8). APPENDIX E, page 191 provides the complete instructions that were taught to the observers. After classroom review and discussion, observers went to a nearby site in the Canyon so that each could practice using the palmtop in accordance with the procedures.

At the suggestion of the TRC, prior to the fieldwork, a binaural recording[40] was made of aircraft sounds in the vicinity of Hanscom Field, Bedford, Massachusetts. Such a recording was hoped to provide a means for training observers who would listen to a playback through headphones, log the sounds heard, and then compare results from one observer to another. During the recording, an observer kept a log of sources that could be compared with the playback results. On playback, the original observer could not repeat the results in the field, and a second subject who was not in the field for the recording, found that binaural information was insufficient. The inability to turn one's head while listening was thought to significantly reduce the ability to identify sources. No further action was taken with the recording.

All teams that went to audibility sites had audiometric testing done to verify that their hearing acuity was within the normal range, as checked by certified audiologists. Each team was provided with field notebooks which contained the material in APPENDIX E, forms for recording latitude and longitude, logging tape recording times, hourly wind speed and direction, and topographic maps of their locations. Global Positioning System (GPS) units were used to help locate sites, and these locations were later crosschecked with topographic maps. Digital watches were provided to each team, and times synchronized prior to departure, and daily by radio time hacks.

presents an example of the resulting observer logs created with this method. At 7:41:43, NPM began logging at Site 4A on 12 September 1999, see Figure 22, page 54, doing a time check at 7:42:50. Logging began at 7:59:59 with a propeller aircraft heard far to the west, and thus could not have been a tour aircraft on the Zuni Point corridor. This prop was heard until 8:02:51 when a jet was heard, flying southwest to northeast. At 8:04:55 only natural sounds were heard – light gusts of wind producing the sound of wind in the foliage. From 8:08:35 until 8:11:38 only air tour helicopters were heard when their sound became inaudible due to the presence of a non-tour propeller aircraft. This information provided the data used to compute the amount of time tour aircraft were audible at each of the audibility logging sites.

A limited check of the consistency of logging procedures was made using data from two sites 5A and 5B where two individuals simultaneously and separately conducted logging. These data, described in Section 8.4.5 and plotted in Figure 40, page 94, were used to estimate measurement error.

[40] Binaural recordings use two microphones, one at each of the ear positions of a dummy head. When played back through headphones, this type of recording is thought to produce an acoustic experience that rivals the realism of the actual sounds.

Table 12. Example Observer Log

NATIONAL PARK SERVICE - GRAND CANYON MODEL VALIDATION - 295870.11					
Site:	4A - Horseshoe Mesa			12-Sep-99	
Time	Acoustic State	A/C Type	A/C Oper	Backgnd Descrip	Comments
7:41:43	Beg Log	***	***	***	Npm
7:42:50	Time Chk	.....	.....	.....	07 42 50
7:59:59	Aircraft	Prop	Other	***	far to west
8:02:06					
8:02:51	Aircraft	Jet	Other	***	sw to ne
8:04:55	Natural	***	***	Wind/Fol	occ lite gusts
8:08:35	Aircraft	Helo	Tour	***	
8:10:14	Aircraft	Helo	Tour	***	second heard before first ended
8:11:38	Aircraft	Prop	Other	***	heard before 2nd tour ended
8:14:50	Natural	***	***	Wind/Fol	
8:15:14	Aircraft	Helo	Tour	***	
8:16:10	Aircraft	Helo	Tour	***	2 helo bef 1st ended
8:18:25	Aircraft	Jet	Other	***	
8:18:54	Natural	***	***	Wind/Ear	
8:23:57					seems to be barely audible stuff, some tonal
8:25:38	Natural	***	***	Wind/Fol	lite gusts from s
8:27:47	Natural	***	***	Wind/Ear	
8:29:59	Aircraft	Jet	Other	***	w to e
8:32:00	Aircraft	Jet	Other	***	2nd jet, far to s
8:33:12	Natural	***	***	Wind/Ear	
8:33:32	Aircraft	Jet	Other	***	same jet to s
8:33:54	Natural	***	***	Wind/Ear	
8:34:01	Aircraft	Prop	Other	***	to se
8:34:47	Natural	***	***	Wind/Fol	
8:35:25	Aircraft	Prop	Tour	***	
8:37:13	Aircraft	Prop	Tour	***	heard bef pref gone
8:40:00	Aircraft	Prop	Other	***	
8:42:25	Aircraft	Jet	Other	***	e to w, s of site
8:43:53	Natural	***	***	Wind/Fol	
8:45:16	Natural	***	***	Insects	when still, hear insect flight
8:46:42	Natural	***	***	Wind/Fol	
8:47:08	Natural	***	***	Birds	can hear aerodynamic sound of swifts
8:47:50	Aircraft	Helo	Tour	***	
8:50:10	Aircraft	Jet	Other	***	
8:50:57	Aircraft	Jet	Other	***	overhead , e to w
8:53:24	Natural	***	***	Wind/Fol	
8:53:39	Aircraft	Prp/Hel	Other	***	to sw

5.2.2 Sound Levels at Audibility Sites

In addition to using the palmtops at all audibility sites, all teams also had digital audio tape recorders, DAT's, with associated microphone and preamplifiers, all battery powered. Digital recordings were made at the sites identified in Table 11, page 55. Standard half-inch microphones were used by four of the teams, while the other four used the low noise system consisting of the Brüel & Kjær (B&K)

Model 4179 1-inch microphone with associated preamplifier and power supply. Five of the teams also used large diameter, two stage, HMMH low noise windscreens[41] (see Table 11, page 55), while the other sites employed the standard B&K Model UA0207 foam windscreen. The tape recorders were run for the full time of the measurements each day, generally from 08:00 to 12:00 and from 13:00 to 17:00, maximum. All systems were calibrated before and after each tape with acoustic calibrators traceable to the United States National Institute of Standards and Technology, NIST.

5.2.3 Sound Levels at Source Site

Though the detailed methods and instrumentation used at the Source Site are reported in a DOT report,[42] this section provides a brief summary of that information. In order to provide high quality reference aircraft source sound levels for all of the computer models, detailed measurements were made of all tour aircraft flying the Zuni Point corridor during the measurement period. The Source Site location was carefully selected under the Zuni Point flight corridor, on the south rim, as far as possible from other noise sources, with a clear view of both helicopter and fixed wing tour aircraft tracks, see Figure 22, and APPENDIX D.

A three-microphone array was used, oriented perpendicular to the south-to-north flight corridor. Spacing from the eastern most to the western most microphones was approximately 1500 feet. Sound level data were fed to a Larson Davis Model 820 sound level meter and recorded digitally on Sony PC208Ax DAT recorders. Aircraft location was recorded with two video tracking systems, one facing east and one facing west, both located somewhat east of the western most microphone. A differential global positioning system (dGPS) precisely determined microphone, and video system component locations, while meteorological data were collected with Qualimetrics Transportable Automated Meteorological Systems, TAMS. During the measurement days, between 08:00 and 12:00 and between 13:00 and 17:00, September 10, 12 and 13, each tour aircraft was identified, logged, and its sound level time history measured and recorded.

5.2.4 Meteorological Data

Meteorological data were collected with Qualimetrics Transportable Automated Meteorological Systems, TAMS, at five sites placed solely during the measurement period, Cedar Mountain, Cremation, Grandview Point, Horseshoe Mesa and Source (for locations, see Figure 22). Data from two permanent stations, Abyss and Hance, were also used.

5.2.5 Air Tour Speeds

Each air tour aircraft was logged by type and time of day at both the Source Site, when the aircraft passed over the centerline of the microphone array, and at Site 5A, Cape Final, when the aircraft passed a clearly identifiable land mark (Gold Hill). These two times were collected on a total of 104 aircraft, and were used to compute average tour aircraft speed for purposes of developing the necessary input for the models. APPENDIX G, Section 6 tabulates these speeds.

[41] For a complete description of the two stage low noise windscreen, see Appendix A of Anderson, *et al*, 1993 listed in footnote 39.

[42] Fleming, *et al*, "Reference Source Data for GCNP Noise Model Validation Study," U.S. Department of Transportation Letter Report DTS-34-FA065-LR2, May 2000.

Page Intentionally Blank

6. DATA REDUCTION

Collected data provided the information for modeling, comparison with results, and diagnostics. Table 13 summarizes the primary types of data and their uses in this study. This section summarizes the reduction of the data to the forms needed here and APPENDIX G, page 213 provides detailed reduction results as used in modeling.

Table 13. Data Derived from Measurements and Their Uses

Type of Data	Uses of Data				
		Modeling		Diagnostics	
	Reduction of Other Data	Run	Computed *V* Measured	Measured Results	Computed *V* Measured
From Audibility Sites:					
Time Lag from Source Site	X				
Audibility Increment	X				
Percent of Time Tours Audible			X	X	X
Various Site Parameters:					
Site Location	X	X			
Site Altitude				X	X
Ambient Sound Level Type	X	X		X	X
Elevation Angle to Corridor				X	X
Perpendicular Distance to Corridor				X	X
Angle of Corridor Visible				X	X
Length of Corridor Visible				X	X
Nearest Met Tower	X				
From Sound Recordings:					
Ambient Sound Levels		X		X	X
Aircraft L_{eq}			X		X
Empirical Detectability Level		(x) [B]		(X) [B]	(X) [B]
Tour Aircraft Speeds [A]	X	X			
From Source Site:					
Tour Operations by Type and Time		X		X	X
Air Tour Sound Levels		X		X	X
Tour Aircraft Speeds [A]	X	X			
From Meteorological Sites:					
Wind Speed				X	X
Wind Direction				X	X
Temperature		X		X	X
Relative Humidity		X		X	X
Barometric Pressure				X	X
Other Data Types					
Topographical Information	X	X			
Flight Corridor Location	X	X			

Notes to Table 13:

[A] Air tour time and location data from audibility site 5A and from the Source Site provided derivation of average tour aircraft speeds.

[B] Uses marked with parentheses (X) are possible future uses, not uses made in the study reported here.

6.1 Audibility Site Data

6.1.1 Time Lag from Source Site

Two of the three computer programs (INM and NODSS) model aircraft operations per time increment, for example, X aircraft flights per hour, and compute results for each site for that hour. In reality, however, for a given hour, not all the sites will experience (hear) the same set of aircraft flights. The sites further along the corridor, such as sites 5 and 9, will hear the tour aircraft at a time later than they are heard at the closer sites, such as sites 4, 6 and 7. In order to make the correct comparison of measured and modeled results for INM and NODSS, the time intervals used to determine the percent of time tours were audible need to be different for each site. For these two models, the tour aircraft as counted at the Source Site per time increment were modeled, and the appropriate time lags for each site were determined by accounting for aircraft speed, and for the time required for sound to travel from the aircraft to each site. (APPENDIX G, Section 9 also describes this issue and the computation of the time difference between the times an aircraft flies over the Source Site and the arrival of the sound at each audibility site.)

To adjust for this time lag between flight over the Source Site, and flight through the audibility range of all audibility sites, a time lag was computed for each site, relative to its distance along the corridor from the source site (to account for aircraft speed) and relative to its distance from the corridor (to account for the speed of sound in air). When the percent time audible values were derived from the measurements for each site, these time lags were used to determine over what time increment the audibility should be determined for proper comparison with computed results.

Table 15 lists the time lags (as well as other parameters discussed below) determined for each site. For example, if the models were to be run for the air tours that flew from 8:00 to 9:00, measured tour audibility at Site 1Ah would be determined from the measurement data for the time increment between 8:02:30 and 9:02:30.

Because the third model, NMSIM, models the tours in the time sequence that they actually flew, for proper comparison of computed and measured results, NMSIM had to compute results at each site using that site's time lag. So, for example, for Site 1Ah, NMSIM computed audibility between 8:02:30 and 9:02:30. Using this offset for NMSIM meant that all four models would be compared with the identical measured results for each site.

6.1.2 Audibility Increment

In order to compute measured audibilities, two questions had to be addressed. First, what time increment should be used (20 minutes, half-hour, full hour, etc.)? Second, for a given site, what amount of the chosen time increment should have been measured for the measured data to be accepted into the analysis?

As discussed above in Section 6.1.1, INM and NODSS compute aircraft noise only for specific time increments. Hence a time increment needed to be selected. In general, the increment needs to be long enough to maximize the likelihood that aircraft that were heard at a site are also included in the modeled numbers, but not so long that information about temporal variability (such as that due to meteorological changes or changes in numbers and types of aircraft) will be lost.

Comparisons of seconds of tour audibility as heard at one of the more distant sites (Site 5A) *versus* number of aircraft counted at the Source Site were made. Figure 24, Figure 25, and Figure 26 show the results. Use of the shorter time increments results in unrealistic matching of tour numbers to

percent of times audible. For example, Figure 24 shows a great many seconds of aircraft audibility during many 20 minute periods, but shows that no aircraft passed over the Source Site during these 20 minute periods. The reason for this miss-match is probably that aircraft are randomly enough spaced and widely enough spaced that the shorter time period is not a representative sample of air traffic. For the match between aircraft heard and aircraft measured to be realistic, the measured audibility during a time increment should be a monotonically increasing function of the number of aircraft observed at the Source Site. Hence the 60-minute period was chosen as the basic analysis increment. All data presented in this report are based on dividing the data collection periods into one-hour increments.

Having selected the one-hour increment, and recognizing that, even if data were collected at all sites for full hour periods, incorporation of the time lag meant that not all one-hour increments for all sites would have a full 60 minutes of audibility logging. Decisions were required to determine whether or not to include in the analysis a site one-hour increment if it were not a full 60-minute observation. For the analysis, "site hour increments" were used if they contained at least 30 minutes of audibility logging. Table 14 shows the distribution of measurements by duration.

Table 14. Distribution of Measurement Durations

Duration of Measurement, Minutes	Number of Measurements
>30=35	6
>35=40	6
>40=45	15
>45=50	6
>50=55	12
>55=60	266
Total	311

Table 15. Average Audibility Results and Various Parameters by Site

Site	Time Lag (mm:ss)	Percent Time Audible			Type of Ambient Environ-ment[A]	Elev. Angle to Corridor (deg)	Perp. Distance to Corridor (mi)	Horizontal Angle of Corridor Visible (deg)	Length Corridor Visible (mi)
		Sep 10	Sep 12	Sep13					
1Ah	2:30	-	1	-	DS	3.9	11.3	22.9	4.7
1Bn	2:19	-	5	-	DS	3.9	11.3	15.5	3.0
2Ah	2:10	1	-	-	W/R3	4.3	9.9	25.5	5.0
2Bn	1:54	-	-	6	W/R2	4.7	9.5	19.1	3.2
2Cn	2:00	1	-	-	W/R4	4.6	9.4	29.7	5.5
2Dh	2:08	-	-	3	W/R4	4.5	9.7	27.6	5.4
3Av	1:08	10	-	-	W/R4	6.1	7.2	44.2	6.0
3Bv	1:00	-	17	-	W/R2	6.2	7.3	31.5	4.2
3Dv	0:51	-	-	5	DS	5.9	7.6	22.8	3.7
3Hn	1:08	9	-	-	W/R3	5.2	7.5	47.9	6.7
3Jn	1:04	-	30	-	W/R3	5.4	7.5	47.8	6.8
3Kn	0:52	-	-	3	DS	5.8	7.7	21.5	3.6
4Ah	0:13	20	29	23	DS	6.2	5.0	67.0	10.5
4Bn	0:21	35	-	-	DS	6.2	5.0	63.1	9.1
4Cn	0:11	-	33	-	PJ	5.6	5.5	61.3	9.5
4Dn	-0:06	-	-	8	PJ	6.0	5.2	29.9	13.9
4En	-0:14	-	-	8	PJ	5.5	5.2	33.4	14.9
5Av	4:03	-	64	49	CF	-0.6	3.7	165.4	25.0
5An	4:03	-	55	48	CF	-0.6	3.7	165.4	25.0
5Bv	4:14	13	-	-	CF	-0.7	4.6	103.2	20.2
5Bn	4:14	13	-	-	CF	-0.7	4.6	103.2	20.2
6Av	-0:01	31	32	26	CF	3.5	1.7	116.2	17.6
6Cn	0:04	-	28	-	PJ	2.8	1.9	118.9	20.5
6Dn	0:18	-	-	28	PJ	2.3	2.1	95.3	12.3
7Ah	2:13	40	-	-	DS	15.4	2.4	107.9	8.0
7Bh	1:42	-	26	-	PJ	10.2	2.3	99.9	23.5
7Ch	1:26	-	-	17	PJ	11.4	2.1	27.3	1.2
7En	2:52	58	-	-	DS	18.7	2.1	151.8	24.3
7Gn	1:52	-	48	-	PJ	11.9	2.1	47.3	21.2
7Hn	1:51	-	-	43	PJ	10.6	2.1	149.3	26.8
8Ah	1:57	14	41	-	PJ	1.4	5.7	103.1	36.2
8Bn	1:27	17	-	-	PJ	2.7	4.0	13.8	1.0
8Cn	1:14	-	29	-	PJ	2.0	4.1	81.9	35.4
8Dh	1:42	-	-	27	PJ	2.5	3.5	109.1	9.7
8En	1:43	-	-	25	PJ	2.4	3.6	106.5	9.0
9Av	12:28	1	-	-	DS	0.6	14.8	43.9	49.7
9Bn	12:12	-	18	-	DS	0.8	11.2	51.3	48.7
9Cv	11:31	-	62	-	W/R1	3.9	2.3	131.2	26.5
9Dv	12:12	-	-	3	DS	0.8	11.2	51.3	48.7
9En	12:12	13	-	-	DS	0.8	11.2	51.2	48.7
9Fn	11:31	30	-	-	W/R1	3.9	2.3	133.6	26.5

Notes to Table 15:

[A] DS = Desert Scrub, PJ = Pinyon-Juniper, CF = Sparse Coniferous Forest, W/RX = Water / Rapids, see text.

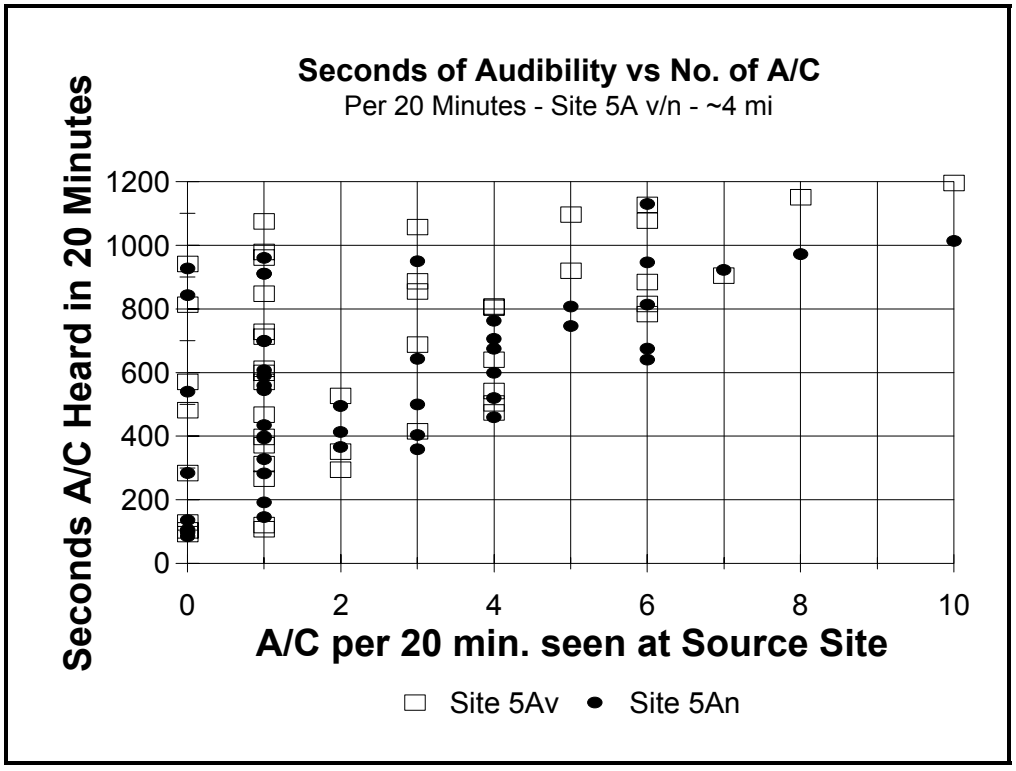


Figure 24. Effect of Using 20-Minute Time Interval

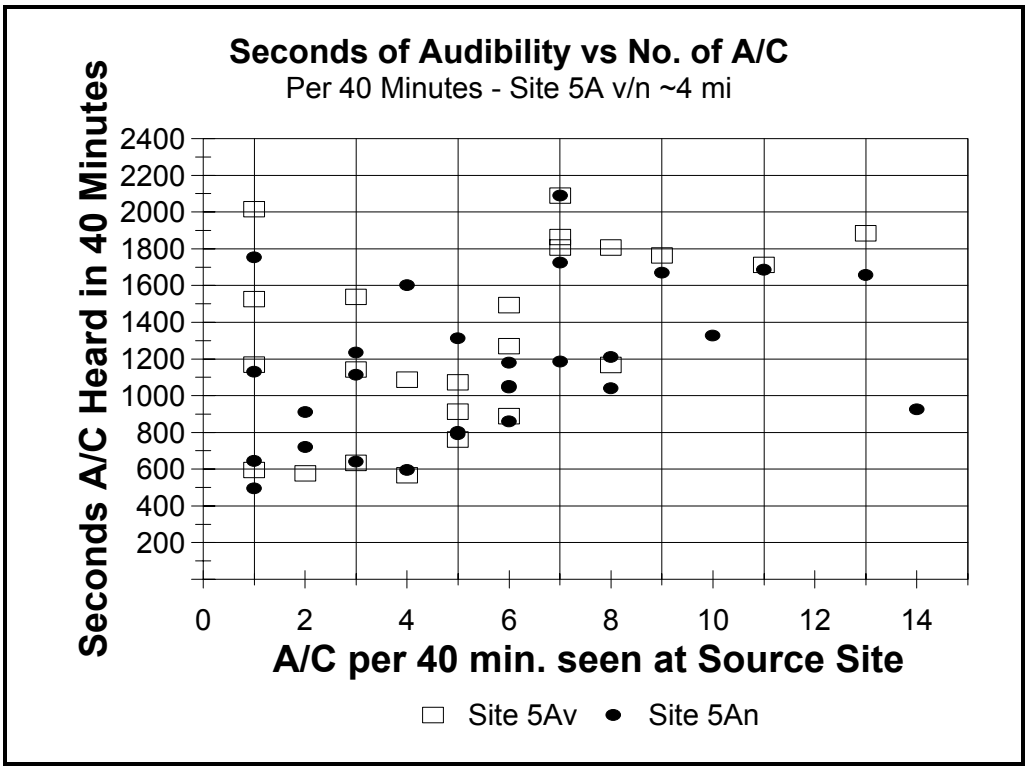


Figure 25. Effect of Using 40-Minute Time Increment

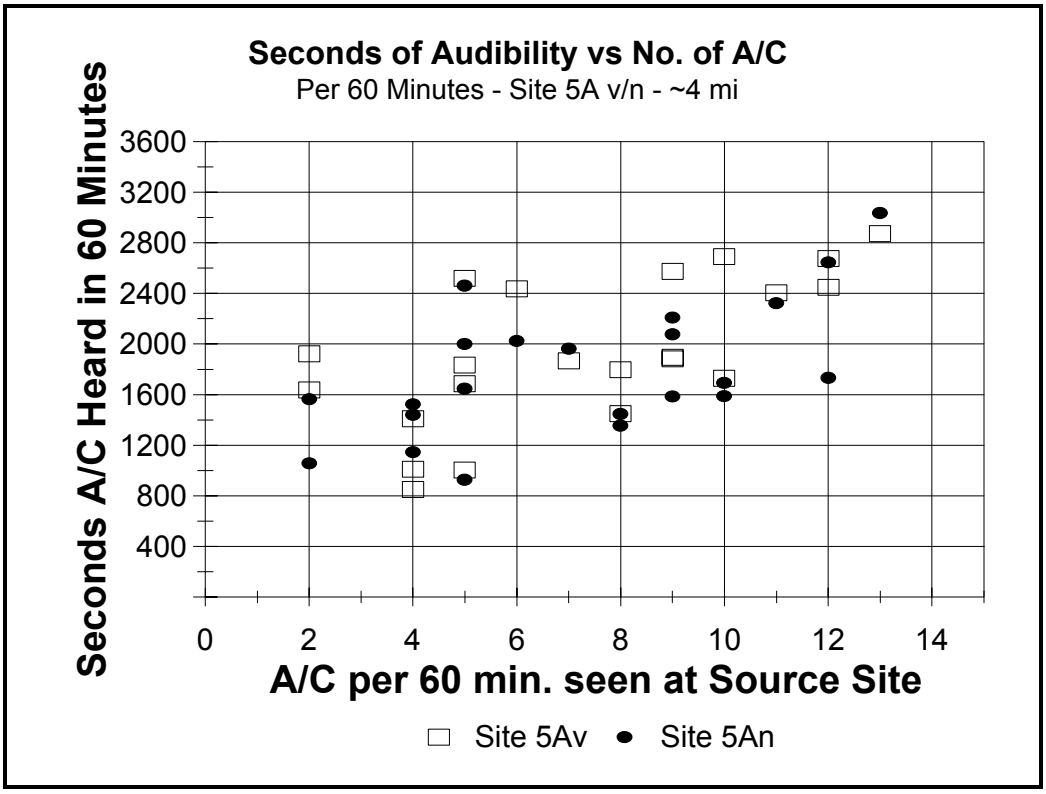


Figure 26. Effect of Using 60-Minute Time Increment

6.1.3 Percent of Time Tours Audible

Having determined the time lags for each site, the time increment and necessary duration per increment for inclusion, the percent of time audible for each site, for each hour increment were computed. Table 15 presents the average results for each site, for each day. These values are the percent of time aircraft were audible for the full day of data collection at each site, and are provided here to convey a sense of the results. The analyses presented in Section 8 used each separate site hour as a data point rather than combining the measured results as done for Table 15. Comparison of the values in Table 15 with site locations, see APPENDIX D or Figure 22, page 54, provides an indication of the relationship between location and audibility duration.

6.1.4 Various Site Parameters

Table 15 also lists primary site parameters used in modeling and analysis.

6.1.4.1 Type of Ambient Environment

The type of ambient environment determined the ambient sound levels used in modeling. For each type of ambient, specific values were assigned, whether A-weighted for the INM 5.1, or spectral (1/3 octave band) for the other models.

For each type of ambient environment, three different values were identified and modeled. First, the levels used for the noise analysis in the FAA's "Environmental Assessment of Special Flight Rules in the Vicinity of Grand Canyon National Park" were modeled. These are the values based on

measurements made at 23 sites in the Grand Canyon during August and September 1992[43] and are referred to here as the "EA ambients." These ambient levels are given in APPENDIX G, Table 11. Second, ambient levels based on the tape recordings made at each site during the data collection step were used to develop the "measured ambients," which were also modeled. (See Section 6.1.5.1 for an overview of their derivation, and APPENDIX F for a more detailed description with resulting ambients.) Finally, as a test of model sensitivity to the ambient sound level assumptions, the measured ambients were increased by 10 dB and modeled.

6.1.4.2 Relationships of Site To Corridor

Several measures of each site's relationship to the corridor were also determined for use in the diagnostic analyses. In acoustic propagation terms, important aspects of the site-to-corridor relationship are the geometric relation of site and corridor as well as any "shielding" of the corridor provided by the Canyon's terrain. Hence, it was thought that parameters that describe these relations and shielding might help in analyzing discrepancies between the computed and measured values. The elevation angle to the corridor is the degrees of elevation that the corridor is above (values greater than 0^o) or below (values less than 0^o) the specific site. Perpendicular distance to the corridor, visible angle of the corridor, and length of visible corridor were also determined, as shown in Table 15, as well as distance to the nearest visible portion of the corridor, and to the center of the visible portion.

6.1.5 Data from Sound Recordings

The DAT recordings made at many of the sites (Table 11) were used to provide three types of quantitative measures for each site. First, the measured ambient was determined for use in modeling. Second, the equivalent level of tour aircraft noise, L_{eq}, was computed for each site, for each hour. These measured equivalent levels are compared with the computed equivalent levels in Section 8. Finally, the value of the detectability level was computed for each site hour, and is provided as information useful for future diagnostics of the models, since those models that use spectral data must assume a detectability level for computation of the time aircraft are audible. The following sub-sections provide summary descriptions of these metrics, their derivation from the DAT recordings and their use in this study.

The tape recordings could be used for determination of each of the measures because each recording could be associated with an observer log made simultaneously. Figure 27 demonstrates how the association of these two types of information facilitated computation of measures of either natural sounds or aircraft sounds by identifying recorded segments of specific sources.

6.1.5.1 Measured Ambient Sound Levels

As discussed in Section 6.1.4.1, three ambients were used in the modeling. The measured ambient levels were determined for each site and time period where tape recordings were made. In general, each tape contained about 4 hours of recorded 1/3-octave band levels. The observer logs determined

[43]See Horonjeff, *et al*, "Acoustic Data Collected at Grand Canyon, Haleakala and Hawaii Volcanoes National Parks," HMMH Report No. 290940.18, NPOA Report No. 93-4, August 1993, and Memorandum to Wes Henry, from Nicholas P. Miller, "Addendum: Natural Ambient Sound Levels for use in Noise Modeling of Grand Canyon NP," February 5, 1999, HMMH Job. No. 295860.05, or National Park Service, "Review of Scientific Basis for Change in Noise Impact Assessment Method Used at Grand Canyon National Park," January 2000.

in which portions of the recordings strictly natural sounds were heard and recorded, and spectral analysis of these portions yielded median values (symbolized as L_{50}) of the natural sound levels for that site, for that four-hour period.

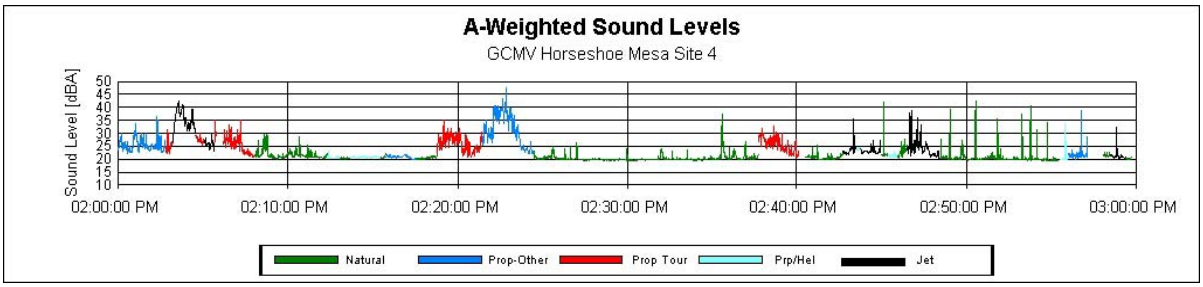


Figure 27. A-weighted Time History Showing Identification of Sources

In many cases, the recorded natural levels were below the level a human with normal hearing could detect. Such levels, if used in modeling, would overstate the audibility of the tour aircraft. Hence, each ambient level was added to the "auditory system noise" that normal humans experience. The result, shown by example in Figure 28, was measured ambient levels that, when compared to aircraft sound levels, would yield realistic computations of the audibility of the sound. For a detailed discussion, see APPENDIX C.[44]

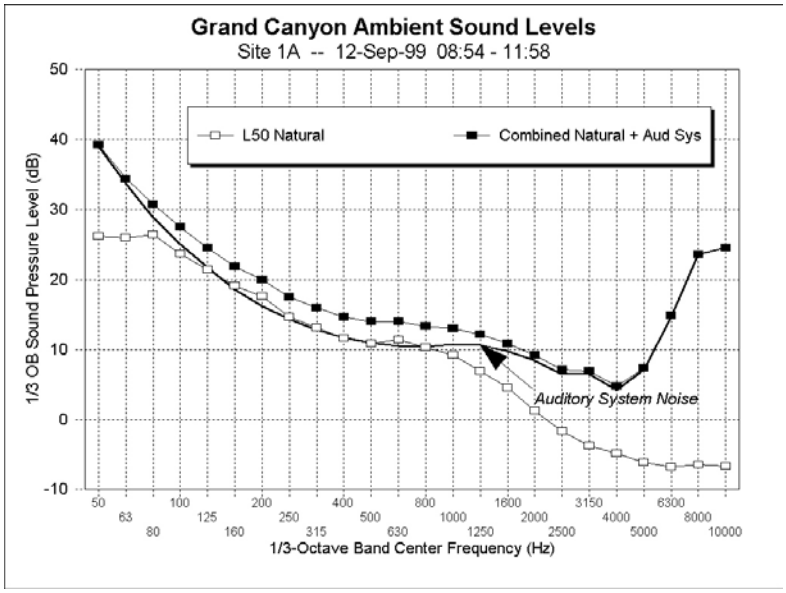


Figure 28. Example Adjustment of Ambient for Human Hearing

[44] Note that for ultimately modeling tour aircraft audibility throughout the Canyon, ambient sound levels need to be generalized to the entire Canyon. The ambient levels measured in this study were examined and combined across similar sites, and were compared with human hearing. Generalized ambient levels for the Canyon were developed, see APPENDIX F.

6.1.5.2 Aircraft Equivalent Levels, L_{eq}

The time periods of the recordings when tour aircraft were audible were identified, and equivalent levels computed for each measurement hour. Table 16 presents the ranges of these equivalent levels for the hours measured, and lists the number of hours measured at each site on each day.[45] Each of the models also computed hourly tour aircraft equivalent levels, and Section 8 presents the statistical comparison of the computed values with the measured values.

Table 16. Range of Equivalent Levels Measured for Each Audibility Site

Site	Range of Tour Aircraft Hourly L_{eq} (dBA)			Number of Hours Measured		
	Sep 10	Sep 12	Sep13	Sep 10	Sep 12	Sep13
1Ah	-	-5 to 23	-	-	6	-
2Ah	0 to 4	-	-	3	-	-
2Dh	-	-	-5 to 16	-	-	7
3Av	0 to 27	-	-	6	-	-
3Bv	-	15 to 21	-	-	8	-
3Dv	-	-	0 to 26	-	-	7
4Ah	22 to 31	5 to 32	14 to 23	7	8	8
5Av	-	22 to 38	21 to 35	-	8	8
5Bv	20 to 30	-	-	8	-	-
6Av	21 to 36	21 to 36	24 to 34	8	8	8
7Ah	29 to 34	-	-	8	-	-
7Bh	-	10 to 23	-	-	8	-
7Ch	-	-	21 to 31	-	-	6
8Ah	0 to 33	16 to 26	-	6	6	-
8Dh	-	-	21 to 28	-	-	3
9Av	0 to 9	-	-	8	-	-
9Cv	-	22 to 35	-	-	6	-
9Dv	-	-	0 to 14	-	-	6

6.1.5.3 Detectability Levels

The DAT recordings together with the audibility logs also provided sufficient information for determining the detectability levels at which observers actually heard air tours. Appendix C.4, page 175, presents the method used and the results. On average, observers first heard tour aircraft (onset of audibility) at a detectability level (10 log (d')) of 5.7 dB, and last heard the aircraft (offset of audibility) at a detectability level of 4.3 dB.

6.2 Source Site Data

Source site data were used strictly for input to the models. These data provided numbers of operations, sound levels by aircraft type, type and time of each aircraft using the corridor and, when combined with the timing of tour aircraft made at Site 5, the speed by aircraft type. APPENDIX G provides these data in detail, Figure 29 summarizes the number of air tours observed flying the

[45] Note that negative values of measured tour aircraft equivalent levels are possible. For example, if one aircraft flight occurred producing a maximum level of 25 dBA and an SEL of 30 dB, the hourly equivalent level would be approximately –5dB.

corridor per hour during the measurement days, and Table 17 gives total numbers of air tours by aircraft type and their average speeds for the three days.

Of primary importance were the tour aircraft sound levels as measured at the source site. These data were developed from the tape recordings and from the aircraft position information collected at the site, see Section 5.2.3. The levels used were those measured at low angles of elevation. Of greatest interest in this study is the ability of the models to compute tour aircraft audibilities at moderate to long distances from the corridor. At these distances, 5 to 10 miles, the angles of elevation from the ground to the aircraft are small and on the order of 1 to 10 degrees, see Table 15, page 64. Because an array of microphones was used to record every tour aircraft flight, the angles of elevation from microphone to aircraft varied for each flight, and it was possible to select and average data for the lower angles. As an example, Figure 30 shows how the sound pressure levels (SPL) varied with angle for the DHC6 (Vistaliner). The values in this figure are corrected to 1000 feet slant distance.

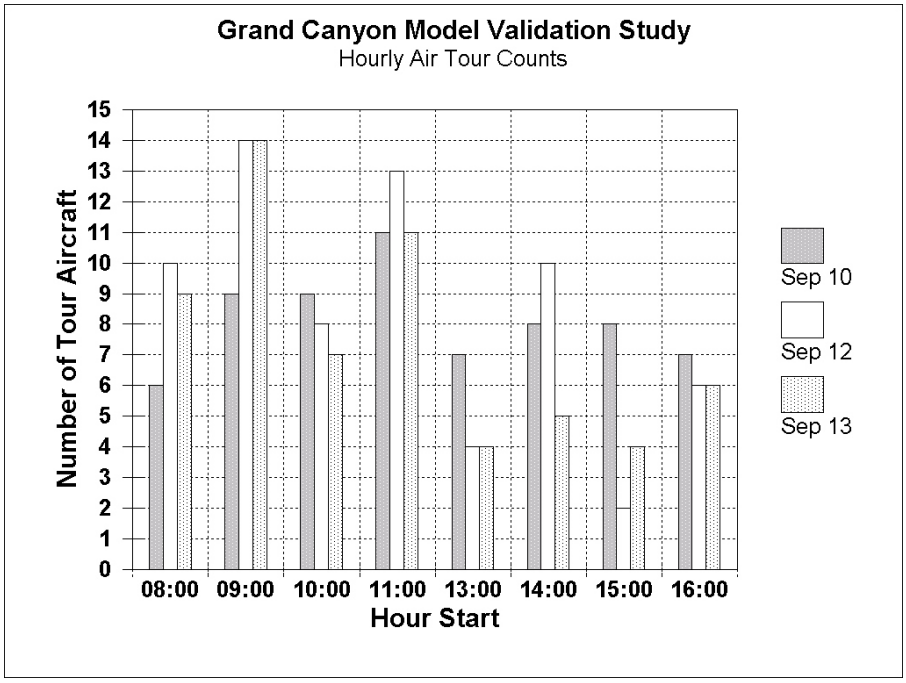


Figure 29. Number of Tour Aircraft Each Hour, Each Day

Table 17. Number and Average Speeds of Air Tour Aircraft

Aircraft Type	Number of Aircraft	Average Speeds (mph)
AS350	22	100
B206B	12	115
B206L	61	108
C207	55	132
C182 (note 1)	4	132
DHC6 (Vistaliner)	38	120
Total	192	-

Note 1. Because of the relatively few measured C182s, its speed is assumed to be the same as that of the C207.

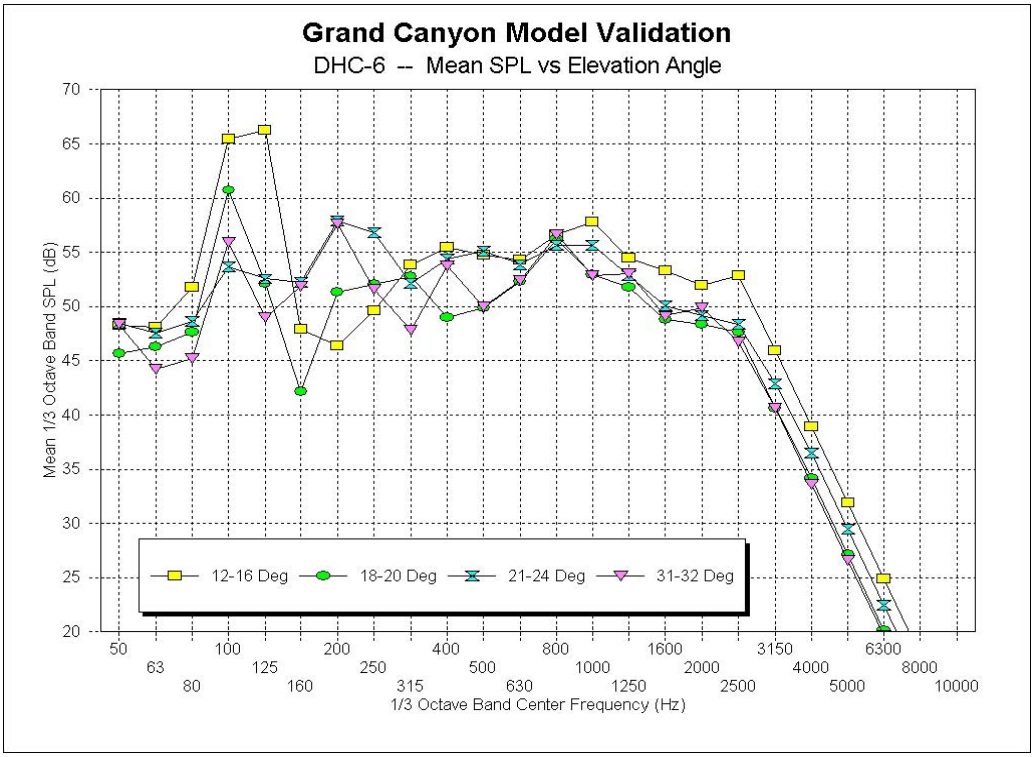


Figure 30. Example of Aircraft Sound Level Variation with Elevation Angle, Corrected to 1000 ft Slant Distance

6.3 Meteorological Site Data

The meteorological data were used primarily for diagnostic analysis of model performance, Section 9. The closest "Met" station was identified for each audibility site and meteorological data from the various stations were then reduced and associated with each measurement site. All data were reduced to hourly values. The primary variables associated with each site were wind speed and direction, temperature, relative humidity and barometric pressure. For each site, the wind component speed in the direction from the nearest visible portion of the track to the site was computed hourly. Because the Met stations were at different elevations (relative to sea level), temperature values from each were used to develop temperature gradients and estimate temperature at each audibility site. Relative humidity and temperature permitted estimation of the atmospheric acoustic absorption for 200Hz and 400Hz per 1000 feet as well.[46]

It should be noted that the meteorological data collected in the Canyon, specifically the wind speed and direction data, were found to be highly variable over short and long distances and over time. Hence, though the conclusions developed from this analysis relative to wind effects reflect the data collected, it cannot be said that these conclusions are definitive for all applications. Conclusions developed from a location where wind speed and direction are fairly constant, both over space and time could well be different.

[46] As discussed in APPENDIX C, detection of aircraft generally occurs at the lower frequencies – 100Hz to 300Hz. Hence, for diagnostic analysis, air absorption at these frequencies is most likely to correlate with aircraft audibility and to affect the relation of computed and measured results.

Page Intentionally Blank

7. MODELING

7.1 Inputs

All inputs used, except the measured ambients, are listed in APPENDIX G; measured ambients are given in APPENDIX C.3.5. Each model used the same inputs, with the following exceptions:

1. INM 5.1 used only A-weighted values of aircraft source levels and ambient sound levels.

2. The Research Version of INM and NODSS used single spectra for each aircraft type and for each ambient.

3. NMSIM used spectra as a function of angle for each aircraft type, and spaced tour aircraft temporally along the corridor as they actually occurred, while the other four models ran total hourly operations.

4. NODSS used an ambient level of –80dBA for calculation of equivalent levels, see Section 3.4.2.

7.1.1 Flight Tracks

Flight tracks were developed and provided by NPS staff working with pilots and FAA. Each model used identical coordinates provided to them separately for the fixed wing and the rotary wing tour corridors as a file attached to the memorandum copied in APPENDIX G, Section 2. Because the validation effort was to be conducted on only the Zuni Point corridor, the western section (the "Dragon Corridor") was not modeled. The entire corridor except for the western leg of the circuit was modeled. The flight tracks modeled ran from the airport in the south, counterclockwise and northward on the east continuing across the northern leg and end at the point where the tracks turn to the southwest to the Dragon Corridor, see Figure 21, page 53 or Figure 22, page 54.

7.1.2 Traffic Counts

Air tour operations as observed at the Source Site, Figure 22, by hour, by aircraft type were provided. NMSIM uses actual times, while the other models use only total hourly traffic.

7.1.3 Audibility Logging Site Locations

Coordinates of the audibility sites, Table 11, page 55 or APPENDIX G, page 213, Table 4, identified the points for which each model computed its results.

7.1.4 Aircraft Sound Levels

Source site measured aircraft sound levels provided each model with the required aircraft source sound levels as the specific, measured speeds.

7.1.5 Ambient Sound Levels

As mentioned, each model was run using three different ambients for each audibility site: the "EA" ambient; the measured ambient; the measured ambient plus ten decibels.

7.1.6 Audibility Log Intervals

Since NMSIM computes sound level time histories, NMSIM was provided with the specific time intervals during which audibility logging was conducted at each site, see APPENDIX G, Section 9.

7.2 Outputs

Each model computed for each site, for each hour interval of measurements, the computed percent of the hour aircraft were audible and the tour aircraft equivalent sound level, L_{eq}. Both INM versions and NODSS models do not have the capability to compute the effect of aircraft sound level "overlap" – the circumstance when a second aircraft becomes audible before the previous one becomes inaudible. In these situations, which are more and more likely to occur as air tour traffic per hour increases, these models will over estimate the percent of time audible. To account for this effect, an empirical "compression" algorithm was determined from audibility data. This compression was derived from data collected in 1992.[47] Numbers of aircraft per time were determined and compared with total number of minutes heard in order to derive a relationship between maximum possible duration and actual duration. These values were fit to an equation, which is presented in Figure 99, page 243, Appendix J.1. (One of the recommendations is that the current study data be used to develop an up-dated compression algorithm that might be applicable to more situations and more parks, see Section 1.11.3.7, page 31.)

[47] Horonjeff, R.D., *et al*, "Aircraft Management Studies, Acoustic Data Collected at Grand Canyon, Haleakala, and Hawaii Volcanoes National Parks," NPOA Report No. 93-4, August 1993.

8. ANALYSIS AND RESULTS

8.1 Overview

Previous sections describe how field data were measured and how corresponding tour-aircraft sound metrics were modeled by computer. This section describes the analysis of these field measurements and model computations, including the results of this analysis.

- Section 8.2 is a tutorial on the *analysis terms* used in this section.
- Section 8.3 restates and expands upon the *study goal* from Section 2, above, and introduces the study's validation matrix.
- Section 8.4 discusses *overall error*—the total discrepancy between model computations and corresponding measurements. Overall error is assessed separately for single-hour computations (rarely needed) and for multi-hour computations (averages over all the study hours at individual sites).
- Section 8.5 discusses *accuracy and bias*, which concern the model's performance on the average—that is, whether the model over-computes or under-computes on the average.
- Section 8.6 discusses *precision and random error*, which concern model performance for every instance of its use—not just its performance on the average. Precision is also assessed separately for single-hour computations and for multi-hour computations.
- Section 8.7 discusses *contour error*, which includes the effects of distance, number of averaged hours, and the computed tour-aircraft sound metric.

Table 18 summarizes the relationship between other report sections and this current section.

Table 18. Relationship Between Other Report Sections and This Current Section

Other report section		Relation to this current section
No.	Title	
5	Data acquisition	Describes acquisition of the measured data that are analyzed in this current section.
6	Data reduction	Describes reduction of the measured data into the format needed for this current section.
7	Modeling	Describes modeling of the computed values that are analyzed in this current section.
9	Insights about model discrepancies	Partially diagnoses the model discrepancies that are plotted in this current section—to provide insight into these discrepancies.
10	Conclusions	Gives conclusions about suitability of the different models for use.
11	Recommendations	Based upon the results in this section and also upon the insights of Section 9, makes three sets of recommendations: (1) application of models, (2) improvements of models, and (3) possible further analyses.
App. H	Model sensitivity to ambient sound levels	Indicates the sensitivity of results in this current section to the ambient sound levels used during computation.
App. I	Further details about model validation: measured versus computed	Supplements this current section with (1) additional graphical comparisons of measured versus computed values, (2) graphical evidence for the validity of hourly averaging, and (3) details about the computation of contour error.
App. J	Further details about measured audibility versus physical factors	Supplements this current section with (1) the method of compressing computed audibilities greater than 100 percent, and (2) some plots of measured audibility versus individual physical factors.
App. K	Input to the non-linear regression	Provides input to the non-linear regression of Section 10.
App. L	The full non-linear regression equation	Derivation of the full non-linear regression equation of Section 10, plus the resulting regression coefficients.

8.2 Tutorial on Analysis Terms

This section defines and discusses the technical terms used in this analysis. The terms occur here in the order they enter the analysis—therefore, in the order they are discussed in later sections.

Figure 31 helps explain the definitions and discussions that follow. The figure uses an illustrative set of data points—a set that exaggerates differences among the technical terms in its panel headings. In this figure, each point is positioned horizontally at a *computed* value and vertically at the corresponding *measured* value. If computations were in perfect agreement with measurements, each point would lie on its plot's diagonal line, from lower left to upper right. Along this line, "measured" equals "computed." Departures from this diagonal line are quantified by the three terms above each panel, which are described next.

8.2.1 Overall error: Single hour (hourly) and multi-hour (site)

Overall error is the total discrepancy between model computations and corresponding measurements. It is the *total* discrepancy because it combines both aspects of model discrepancy—model accuracy (Section 8.2.2) and model precision (Section 8.2.3).

In the upper-left panel of Figure 31, overall error is small because nearly all the points lie close to the diagonal—that is, all their vertical distances from the diagonal line are small. The figure's other panels show three reasons why overall error can be large:

- **Upper-right panel:** Overall error is moderate because of lower precision. The *scatter* from point to point is large, even though the average match with the diagonal is close. Low-precision models have large random errors (scatter). This type of overall error is generally preferable to that shown in the lower panels; this type of error can be accommodated by repeated trials of the model in an appropriate manner since, on average, the results agree with the measurements.
- **Lower-left panel:** Overall error is large because of low accuracy. *On the average*, the points lie far away from the diagonal. Low-accuracy models have large bias errors and this type of error is difficult to accommodate; repeated trials can be expected to always give biased results.
- **Lower-right panel:** Overall error is large for both of these reasons.

In brief, small overall error requires high accuracy (close match on the average) and also high precision (close match for every single point—that is, low scatter about the average). Because overall error accounts for both these aspects of "measured versus computed," it is used as the primary metric for model validation in this study. Two overall errors result for each computer model:

- *Single-hour error (also called hourly error): Overall error for single-hour computations at individual sites*. Single-hour error results when the "measured versus computed" plots contain a point for each measured site-hour. Single-hour error is relevant only rarely—when tour-aircraft sound is computed for an individual hour at a specific site.
- *Multi-hour error (also called site error): Overall error for multi-hour computations at individual sites or at site groups*. Multi-hour error results when the "measured versus computed" plots contain only one point for each site or site group—a point that is the average of all study hours at that site or at a group of sites. Multi-hour error is relevant for most computer-model use—when tour-aircraft sound is computed for individual sites (averaged over many hours of tour operation) or for sound contours over all possible sites.

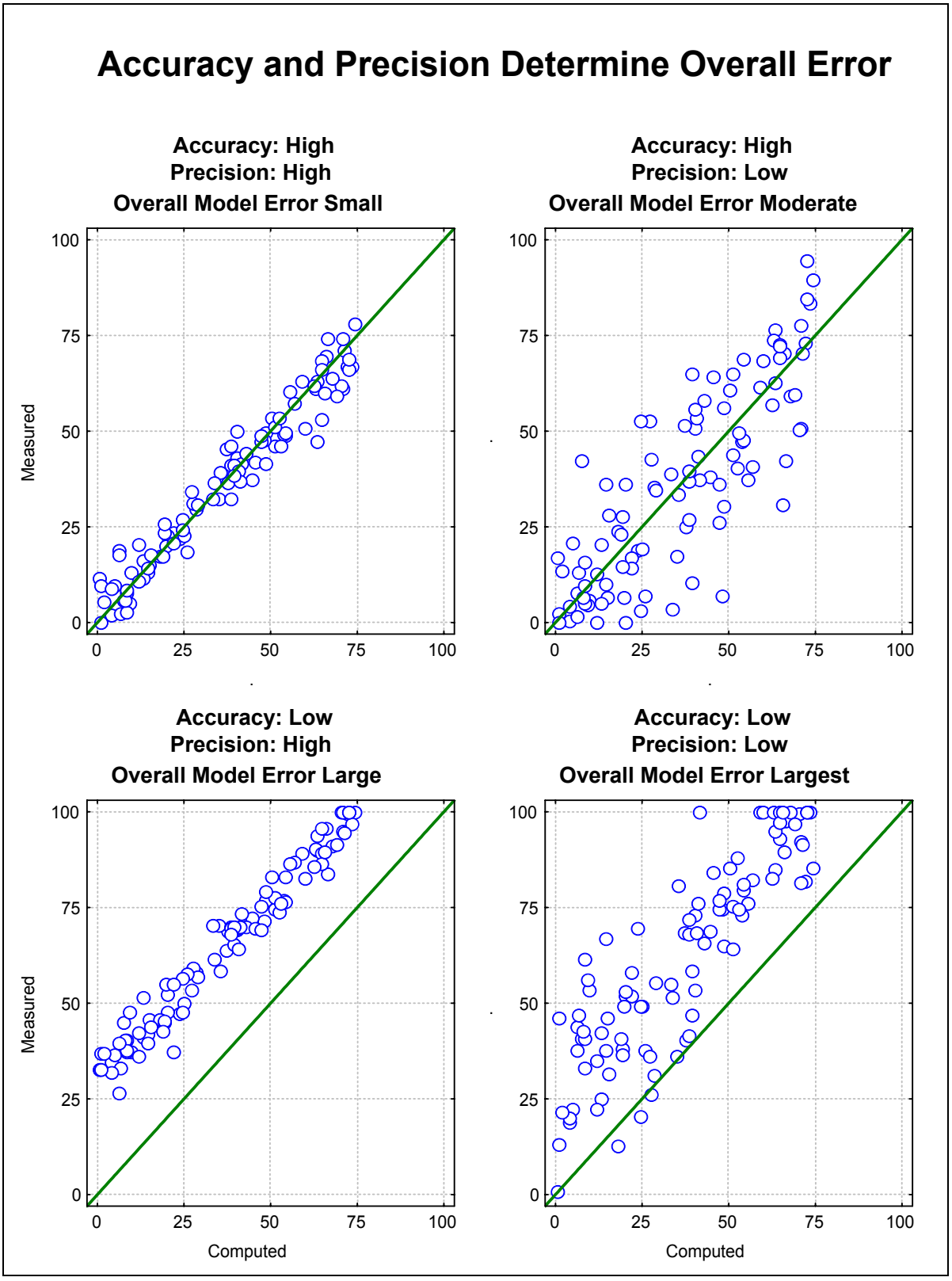


Figure 31. Illustration of Accuracy, Precision, and Overall Model Error

8.2.2 Model accuracy (bias portion of overall error)

Model accuracy is a measure of a model's performance on the average—that is, how well average computations match average measurements. If the average of all measured values is subtracted from the average of all computed values, the resulting single-number "bias" quantifies the model's accuracy. If this bias is nearly zero, the model is judged accurate. Otherwise it is less so, depending upon the magnitude of the bias.

In addition, this single-number bias can be diagnosed as a function of other factors. Figure 31, above, shows one such diagnosis—bias as a function of computed value. In this diagnosis, accuracy is high (and independent of computed value) in the two top panels, because an "average-trend line" through their points would closely match each panel's diagonal line. This tight average match shows *high accuracy* and therefore *small bias error*. More specifically, it shows that accuracy is high (bias nearly zero) over the full range of computed values, from left to right in the graph.

In contrast, accuracy is low in the two bottom panels of the figure, because of the mismatch between such an average-trend line and the diagonal. This mismatch, on the average, shows *low accuracy* and therefore *large bias error*. These panels show low accuracy over the full range of computed values.

Similar plots diagnose accuracy (bias) as a function of various physical factors, such as visible angle of the flight track. Many such plots appear in later sections. They are useful to learn whether or not model bias depends upon these physical factors. When it does, a model may be of questionable use for some numeric ranges of these physical factors.

8.2.2.1 Diagnosis of bias as a function of computed level

For more complex data sets, accuracy is somewhat more complex than shown in Figure 31, however. Figure 32 illustrates this additional complexity. Figure 32 contains a diagonal line and a set of "measured versus computed" data points, coded by the study's site-groups. This coding is not important here.

In addition, the figure contains three curved lines—a central one bounded by two flanking ones. The central curved line is a *regression line*—a special average-trend line—through the figure's points. This regression line shows the overall trend of the data points, as a function of computed values.

The flanking curves are 95-percent *confidence bounds* on this regression. These confidence bounds are needed because sites and measurement hours were "sampled" for this study. Every possible hour at every possible site was not measured and computed. Because of sampling, we are not perfectly confident of the average relationship between measurements

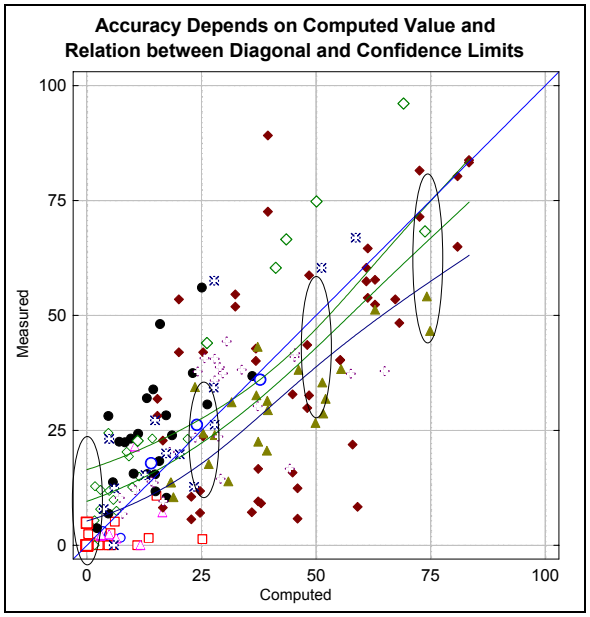


Figure 32. Illustration of Bias Diagnosis with Computed Value

and computations—that is, of the vertical position of the regression line. At best, we are 95-percent confident that the actual vertical average (all possible hours at all possible sites) lies somewhere between these two confidence-bound curves.

8.2.2.2 Quantification

Using Figure 32, accuracy is diagnosed (as a function of computed value) at four horizontal locations, shown by ellipses in the figure:

- *Ellipse at 0:* Within the ellipse at zero, the regression line's upper and lower confidence bounds both lie *above* the diagonal—by vertical distances of 16 and 6, respectively. In words, computations near zero are *lower* than their corresponding measurements. Numerically, these computations have *negative bias*—somewhere between –16 and –6 on this graph. The numerical values are negative because bias is always "computed minus measured."

  This *range* of possible bias (–16 to –6) is the quantitative measure of model accuracy, as a function of computed value. Both these bias values have the same algebraic sign, because the diagonal line lies outside the confidence region. Because they have the same sign, we are 95-percent confident of model bias in this region—that is, we are 95-percent confident that the measured bias (–16 to –6) is not due to random error.

  In this region of computed values, the model has low accuracy. It has negative bias perhaps as large as –16. Computations around zero occur at extremely large distances, however, so perhaps this bias is not too important to a model's use.

- *Ellipse at 25:* Within the ellipse at 25, one confidence bound lies above the diagonal, while the other lies below it—by vertical distances of 4 and 8, respectively. In words, measurements in this region might be lower or might be higher than computations, on the average. Numerically, computations are biased in this region between –4 and +8. Because the diagonal line lies vertically between the two confidence bounds here, we are not 95-percent confidant that the model is truly biased in this region, even though the diagonal lies slightly above the regression line. This small offset is likely the result of random sampling error.

  In this region of computed values, the model is relatively accurate. It is not necessarily biased, and the largest potential bias is only 8.

- *Ellipse at 50:* Within the ellipse at 50, both confidence bounds lie below the diagonal line—that is, measurements are lower than computations. Numerically, computations are biased in this region between +2 and +12. As a result, we are 95-percent confident of model bias in this region, because the diagonal line lies vertically outside the two confidence bounds.

  In this region of computed values, the model has lower accuracy. It has positive bias possibly as large as 12. However, this maximum value is not particularly large, so the model's inaccuracy may not be important in this region.

- *Ellipse at 75:* Within the ellipse at 75, the upper confidence bound lies just on the diagonal. In this region, computations are biased between 0 and +18. As a result, we are not 95-percent confidant that the model is biased in this region, even though the diagonal lies above the regression line. The moderately large offset is likely the result of random sampling error.

  In this region of computed values, the model is relatively accurate. It is not necessarily biased, although the largest potential bias is moderately large—a value of 18.

In all, a model may have a small single-number bias—as in this example—but may be significantly biased in different regions of computed values. This happens when one region's positive bias offsets another region's negative bias in the computation of the single-number bias. Diagnosis of bias, as in this example, helps assess model usefulness in different regions of computed values.

8.2.2.3 Unknown accuracy

The ellipse at 75 in Figure 32 hints at another possibility, but does not show it for this data set. For a different data set, what if (1) the upper confidence bound remained the same as in the figure, but (2) the lower bound moved much further downward? This would happen if the data set had much more scatter than shown in this figure.

In this modified example, again we would not be 95-percent confidant that the model is biased. However, the potential bias would be very large—say, 50. In this situation, the model's accuracy is simply not known with any reasonable confidence. The range of possible bias is too large (0 to 50). The data are scattered too much to determine model accuracy in this instance.

This discussion of model accuracy and bias will be useful in understanding Section 8.5, below.

8.2.3 Model precision (random portion of overall error)

As discussed above, model accuracy measures a model's average performance—its bias. Averages over many hours and many sites are not always appropriate, however. Sometimes computations are needed for an individual hour, or for an individual site. Whenever averages are not sufficient, then model precision is important.

Model precision is a measure of the model's random error. The smaller this random error, the better is the model's precision. Model precision says how well model computations match measurements—for each measured hour, or for each measured site, whichever is relevant. A model is more precise if it closely matches every single measurement, rather than just the average—that is, if its computations have high correlation with corresponding measurements.

Model precision is high in the two left panels of Figure 31, above. In those two panels, correlation between computations and measurements is high. Note that this correlation does not require the average match to be close, as well. It just requires tight clustering of vertical values at every location along the horizontal axis.

In this study, two precision values result for each model:

- *Single-hour precision (also called hourly precision)* results when the "measured versus computed" plots contain a point for each measured site-hour. Single-hour precision is relevant only rarely—when tour-aircraft sound is computed for an individual hour at a specific site.
- *Multi-hour precision (also called site precision)* results when the "measured versus computed" plots contain only one point for each site or site group—a point that is the average of all study hours at that site or site group. Multi-hour precision is relevant for most computer-model use—when tour-aircraft sound is computed for individual sites (averaged over many hours of tour operation).

This discussion of model precision and random error will be useful in understanding Section 8.6, below.

8.2.4 Contour error: Effect of distance, number of averaged hours, and computed metric

For this analysis, the analysis of contour error has produced estimates of the 95% confidence limits for tour-aircraft sound contours. In the analysis described in Section 8.7, contour error is by a re-analysis of model error as a function of:

- Distance from the flight track,
- Number of hours averaged during tour-aircraft sound computations, and
- Computed tour-aircraft percent time audible or equivalent level.

With this additional analysis, overall error can be estimated anywhere on contours computed by the study's models.

8.3 The Study Goal: Restatement and Expansion

This study's analysis compares "measured versus computed" tour-aircraft sound, in order to:

Determine the degrees of accuracy and precision that existing computer models provide, in comparison with field measurements, in the calculation of the percent of time tour aircraft are audible in the Canyon, and calibrate one or more of these models to provide a tool for computation of air tour audibility in the Canyon.[48]

In tabular form, the study goal is to fill in the empty cells of Table 19, the study's validation matrix—based upon the computations and measurements reported in previous sections of this report. The following sections review the meaning of entries in this table.

Table 19. Validation Matrix, but Without Validation Results

Metric	Ambient sound levels used in computation	Computer model	Number of site-hours	Components of model validation			
				Overall error	Accuracy	Precision	Contour error
Audibility	Measured	INM (A levels)	192				
		INM (1/3 octaves)	192				
		NMSIM	192				
		NODSS	192				
	EA	INM (A levels)	301				
		INM (1/3 octaves)	301				
		NMSIM	301				
		NODSS	301				
Equivalent Level (L_{eq})	————	INM (A levels)	147				
		INM (1/3 octaves)	147				
		NMSIM	147				
		NODSS	147				

[48] In addition to examining the "percent of time audible", the tour aircraft "hourly equivalent sound level," L_{eq} was also examined. This equivalent sound level is a measure of the total sound energy produced by tour aircraft during an hour, and is the metric commonly used in Environmental Assessments, Environmental Impact Statements and other common types of environmental analyses.

8.3.1 *Two metrics of tour-aircraft sound, combined with ambient sound levels used in computation*

As shown in Table 19, each computer model is assessed for two metrics of tour-aircraft sound:

- Tour-aircraft audibility, computed with:
  - Measured-ambient sound levels, and
  - EA-ambient sound levels.
- Tour-aircraft equivalent level (L_{eq}), which does not depend upon ambient sound levels in its computation.

8.3.2 *Four computer models*

As shown in Table 19, model computations are assessed for four existing computer models:

- INM (A levels)—the Integrated Noise Model (INM), version 5.1, which does its computations using only A-weighted levels,
- INM (1/3 octaves)—the INM in its Research Version, which includes one-third octave band spectral information. Both INM models, which are energy based, account for differences in site elevation, but not for shielding due to terrain,
- NMSIM—the NOISEMAP Simulation Model (NMSIM), version 2.3a, which also uses spectral information, accounts for park terrain, computes tour aircraft audibility, flies aircraft in the actual time sequence in which they occurred, and includes the directivity of each aircraft type, and
- NODSS—The National Park Service Overflight Decision Support System (NODSS), which uses spectral information and was designed to account for park terrain features, and to compute tour aircraft audibility.

8.3.3 Number of site-hours

As shown in Table 19, the number of site-hours differs for the three major portions of the table. Figure 33 shows all combinations of sites and hours that were measured in this study. Sites appear vertically, while days and hours stretch along the bottom. For example, Site 1Ah was measured on 12 September, during the six hours shown in the figure.

In all, 301 site-hours were measured. This full set of site-hours was used to validate audibility when it was computed with EA-ambient sound levels. However, ambient sound levels were *measured* at only 192 site-hours (from digital tape recordings). For this reason, only 192 site-hours could be used to validate audibility when it was computed with measured-ambient sound levels.

In addition, tour-aircraft sound levels were loud enough to be measured at only 147 of these tape-recorded site-hours. Therefore, validation of equivalent level is restricted to these 147 site-hours. This number falls short of the full 192 hours of tape recording, because some hours had audible aircraft with equivalent levels too low to accurately measure from the tape recordings. For such hours, aircraft were heard and aircraft equivalent levels were computed by the computer models, but measured aircraft equivalent levels could not be determined.

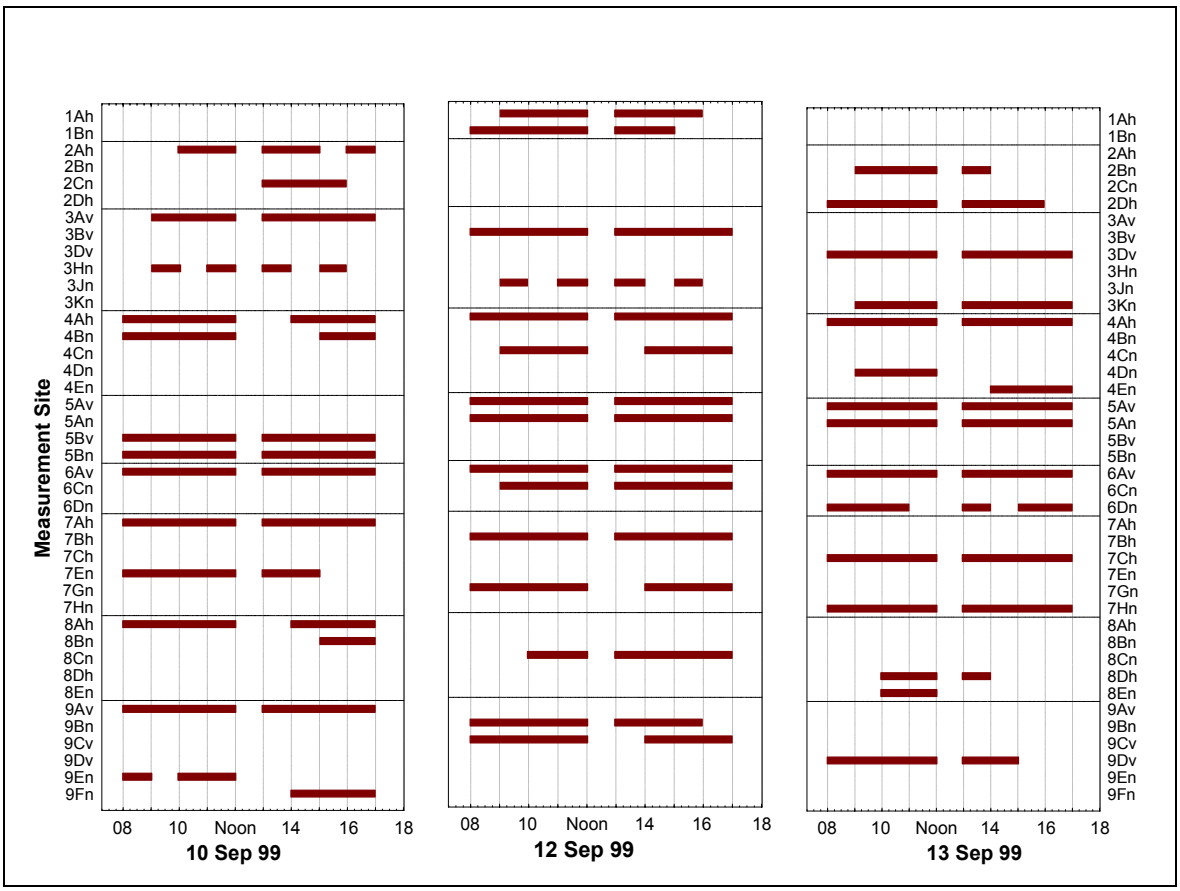


Figure 33. All Combinations of Sites and Hours Measured in This Study

8.3.4 *Components of model assessment*

As shown in Table 19, model assessment combines:

- Overall error,
- Accuracy,
- Precision, and
- Contour error.

8.4 Overall Error

8.4.1 Overview

This section discusses overall error. This overall error is the total discrepancy between model computations and corresponding measurements. It is the *total* discrepancy because it combines both component aspects of model discrepancy—model accuracy (Section 8.5) and model precision (Section 8.6).

First, this section graphically compares "measured versus computed" sound from tour aircraft. From this comparison, it then computes each model's overall error. Two overall errors result for each computer model:

- *Single-hour error (also called hourly error): Overall error for single-hour computations at individual sites*. Single-hour error results when the "measured versus computed" plots contain a point for each measured site-hour. Single-hour error is often relatively large, because the computer models cannot generally account for specific hourly meteorological conditions (or other factors that cause hour-to-hour variability). However, single-hour error is relevant only rarely—when tour-aircraft sound is computed for an individual hour at a specific site.
- *Multi-hour error (also called site error): Overall error for multi-hour computations at individual sites*. Multi-hour error results when the "measured versus computed" plots contain only one point for each site-group—a point that is the average of all study hours at that site-group. Multi-hour error is relevant for most computer-model use—when tour-aircraft sound is computed for individual sites (averaged over many hours of tour operation) or for sound contours over all possible sites. Multi-hour error is generally smaller than single-hour error, because hour-to-hour changes in meteorology—plus other hour-to-hour causes of model discrepancy—tend to balance out over time.

8.4.2 Mathematical computation

Mathematically, overall error is the "root-mean-square" (rms) difference between computations and measurements—that is, between each computed value and its corresponding measured value:

$$\left(\begin{matrix}\text{Overall}\\ \text{error}\end{matrix}\right) = \sqrt{\frac{1}{P}\sum_{i=1}^{P}\left(S_{i,\,computed} - S_{i,\,measured}\right)^2}. \qquad (1)$$

Under the square root sign, the computed and measured S_i stand for the tour-aircraft sound metric of the i^{th} point—either audibility or equivalent level—and P is the number of points in the computation. For single-hour (hourly) error, all site-hours are included in this computation. For multi-hour (site) error, only one point is included per site-group—the average of all that site-group's hourly values.

8.4.3 Single-hour (hourly) overall error

For each site-hour in Figure 33 above, analysis starts with a computed value and a measured value of tour-aircraft sound—a pair of numbers that can be plotted against each other for comparison. Figure 34 through Figure 36 contain these plots, with the points labeled by "site group". (Table 20 below identifies which sites are in which groups.) These figures graphically assess "measured versus computed" values, for both audibility and equivalent level, for all four models. Two figures are required to assess audibility, because it was computed with two different sets of ambient sound levels—measured ambient and EA ambient. Appendix I.1 contains several supplemental plots.

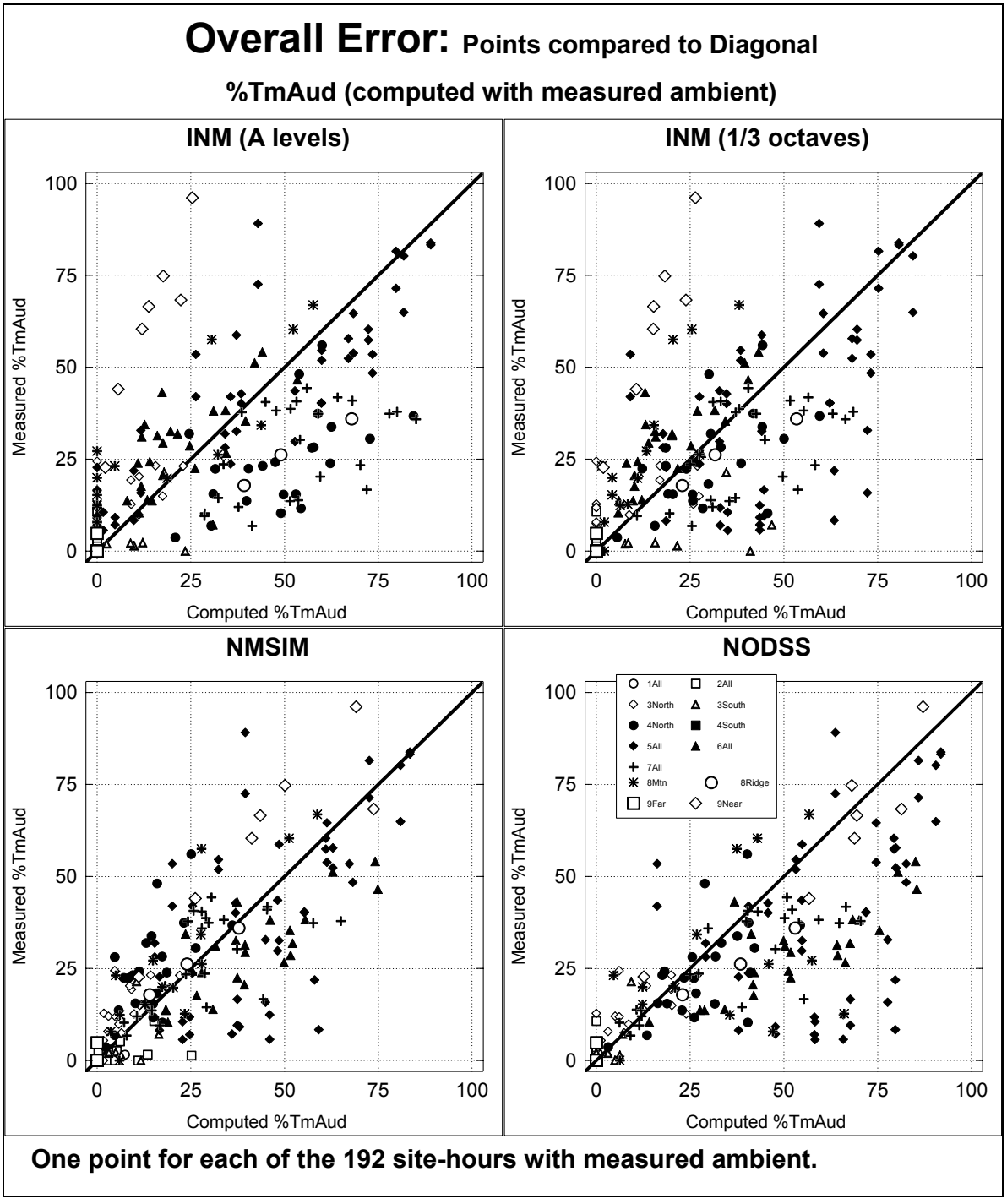


Figure 34. Single-Hour Overall Error: Audibility, Computed With Measured Ambient

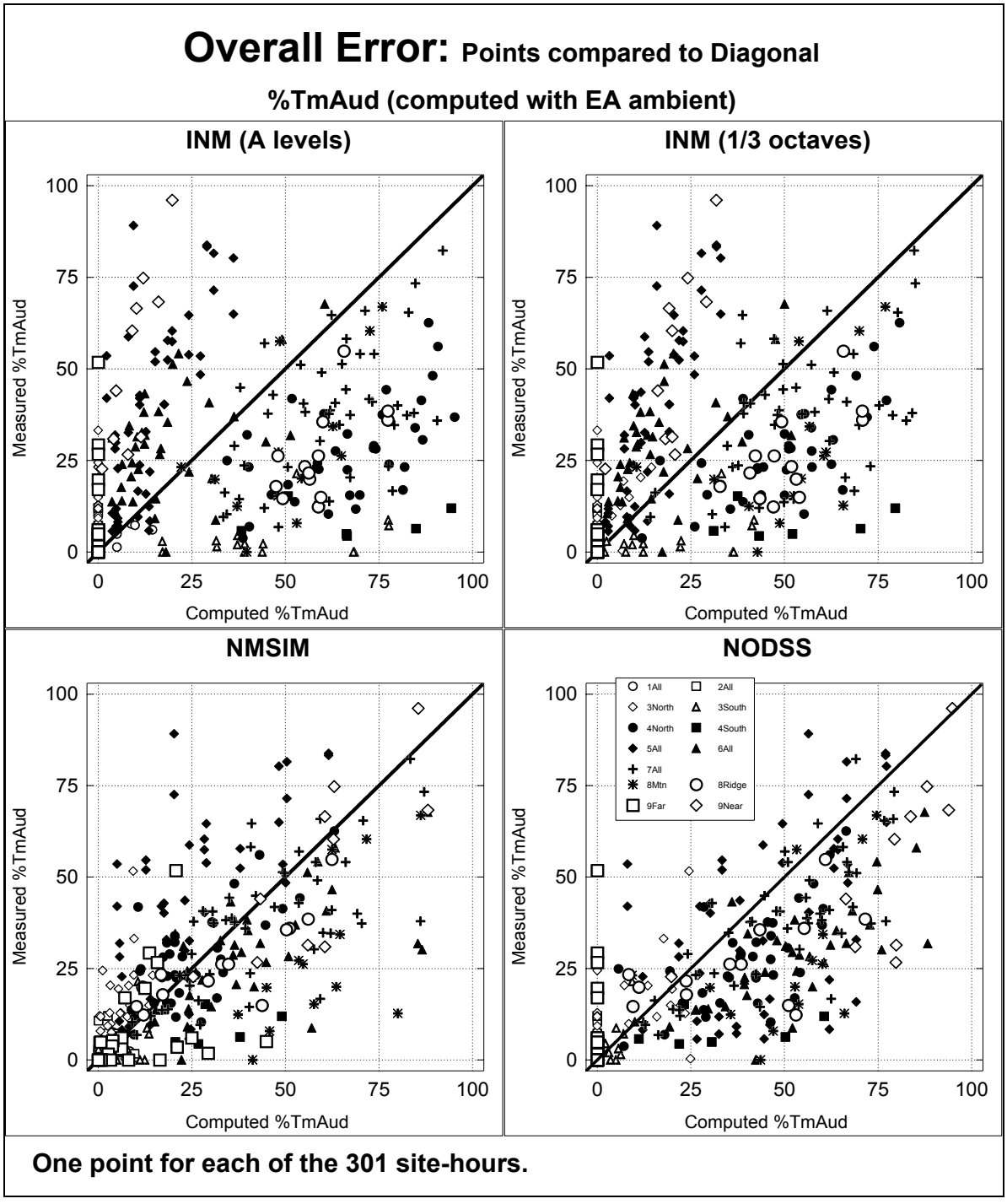


Figure 35. Single-Hour Overall Error: Audibility, Computed With EA Ambient

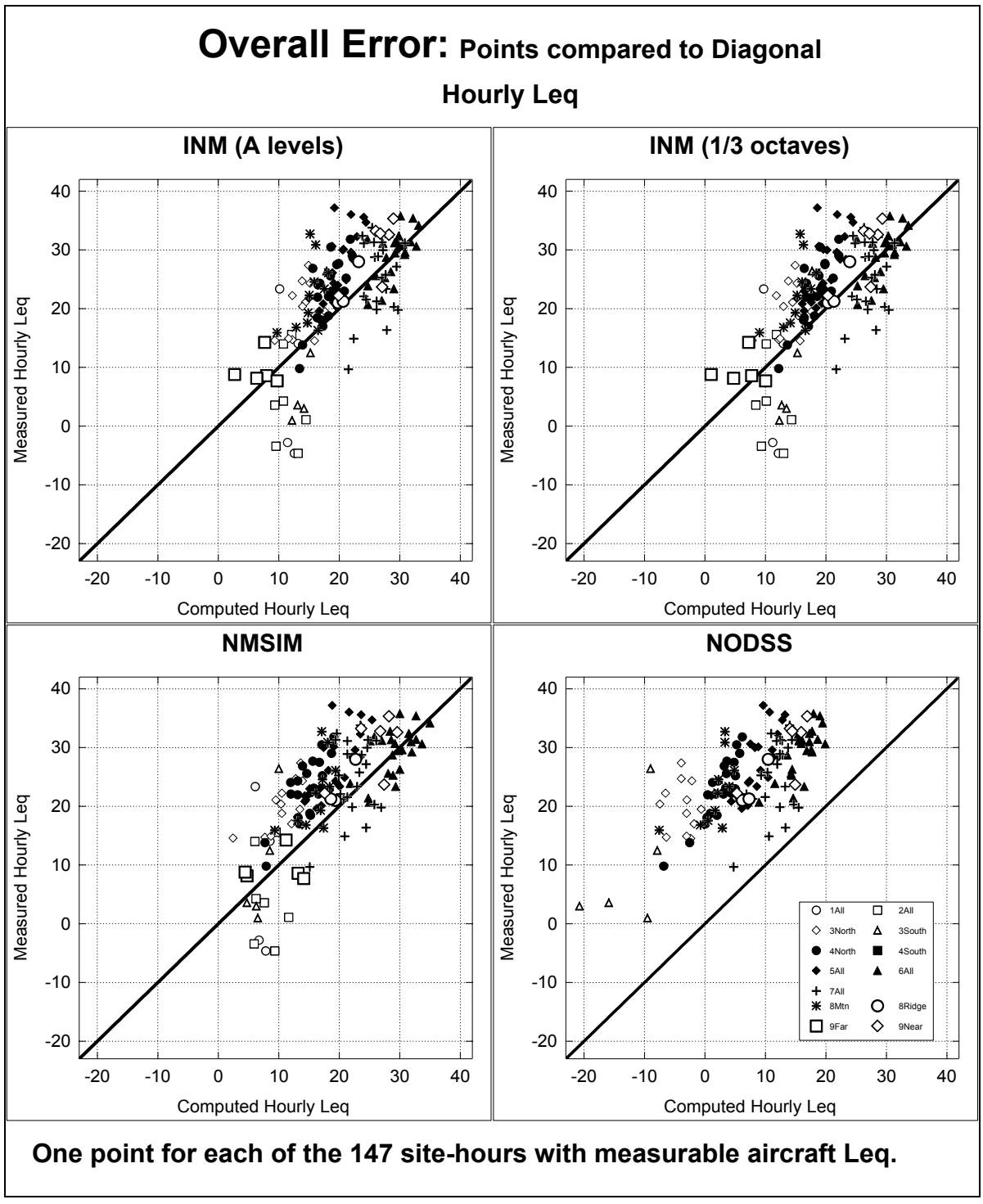


Figure 36. Single-Hour Overall Error: Equivalent Level

Figure 36 contains negative sound levels, which are rarely encountered. They are not impossible, however, as might be thought. They occur when aircraft sound is audible and measurable, but for only a brief portion of the hour and when non-aircraft sound levels are low – making low level aircraft sound measurable. The equivalent level is a level that is "spread" over the time period in question. Hence, for low level, short-duration events, a negative value is possible when adjusted for the longer time period.

In these three figures, each point is for one specific site-hour. It is positioned horizontally at the *computed* value for that site-hour, and vertically at the corresponding *measured* value. If computations were in perfect agreement with measurements, each point in these three figures would lie on its plot's diagonal line, from lower left to upper right. Along this line, "computed" equals "measured." As is obvious from these figures, measured values scatter up and down from this diagonal line. Overall error is low when points gather closely around the diagonal. It is high when they scatter significantly—or when they gather closely together, but offset from the diagonal, as when there is bias (precision) without accuracy (recall Figure 31).

The coded points in the three previous figures show the resulting clustering of the data, group by group. Table 20 shows group membership. See the site figures in Appendix D for further detail.

Table 20. Site-Group Membership

Site-group	Group membership
1All	1A, 1B
2All	2A, 2B, 2C, 2D
3North	3A, 3B, 3H, 3J
3South	3D, 3K
4North	4A, 4B, 4C
4South	4D, 4E
5All	5Av, 5An, 5Bv, 5Bn
6All	6A, 6C, 6D
7All	7A, 7B, 7C, 7E, 7G, 7H
8Mtn	8A
8Ridge	8B, 8C, 8D, 8E
9Far	9A, 9B, 9D, 9E
9Near	9C, 9F

Table 21 contains the resulting single-hour overall errors from these three figures, separately by the type of sound metric, by ambient sound levels used in computation, and by computer model. The table's fourth column is computed from the points in these figures, using Eq. (1), above. Each model's computed values should be reported as "the computed value, plus/minus these tabulated overall errors." Note that NODSS' equivalent level computations show a distinct, significant bias. The source of this bias may be due to the way in which NODSS had to be used to make these computations, (described in Section 3.4.2), but its specific cause is unknown.

Table 21 is relevant only on those rare occasions when a model is needed to compute single-hour values at specific sites, rather than multi-hour values.

Table 21. Validation Matrix: Single-hour (Hourly) Overall Error

Metric	Ambient sound levels used in computation	Computer model	Overall (rms) error
Audibility	Measured	INM (A levels)	20 %TmAud
		INM (1/3 octaves)	19 %TmAud
		NMSIM	14 %TmAud
		NODSS	22 %TmAud
	EA	INM (A levels)	30 %TmAud
		INM (1/3 octaves)	24 %TmAud
		NMSIM	17 %TmAud
		NODSS	20 %TmAud
Equivalent Level (L_{eq})	————	INM (A levels)	7 dB
		INM (1/3 octaves)	7 dB
		NMSIM	8 dB
		NODSS	18 dB

8.4.4 Multi-hour (site-group) overall error

As shown in Figure 33, above, field measurements were made at a total of 41 sites, labeled 1A through 9F.[49] At some of these sites, measurements were made over many hours—a maximum of 24 hours at site 6A. In contrast, at other sites measurements were made for only a few hours—a minimum of 2 hours at sites 8B and 8E.

This disparity in measured hours, from site to site, complicates the analysis of multi-hour (site) overall error. To compensate, nearby sites were combined in the analysis into a total of 13 "site-groups." This grouping balanced out the numbers of measured hours, which is preferable for analysis. Only geographically nearby sites were grouped together, and then only those with similar views of the tour-aircraft flight corridor.

To graphically assess multi-hour (site) overall error, Figure 37 through Figure 39 average all hours in the previous figures into one point per site-group. Comparison of these three figures with Figure 34 through Figure 36 shows that this averaging reduces the scatter of the points about the diagonal lines in several instances, whereas it does not reduce it as much in other instances. For example, comparison of Figure 34 with Figure 37 shows that averaging over many hours reduces scatter for all models, more for NMSIM and NODSS, and not as much for the INM versions.[50]

Table 22 contains the resulting multi-hour (site) overall errors from Figure 37 through Figure 39, separately by the type of sound metric, the ambient sound levels used in computation, and by computer model. The table's fourth column is computed from the data in these plots, using Eq. (1)

[49] Note that two sites—5A and 5B—appear twice in the figure, since two independent teams measured them.

[50] Note that more measured values are averaged when using the EA ambient than when using the measured ambient. The measured ambient results, Figure 37, could be computed for only those measurements where tape recordings were made, while the EA ambient results, Figure 38, could be computed for all sites. Thus, some points have different measured values in these two figures. For example, "9Far" sites were measured at about 10% TmAud when all 9Far sites were included, Figure 38, but close to 0% TmAud when only the 9Far sites with tape recordings were included, Figure 37.

on page 84 and gives the magnitudes of the rms error about the diagonal. Table 22 is relevant when a model is used to compute multi-hour sound metrics—that is, averages over many hours at individual sites.

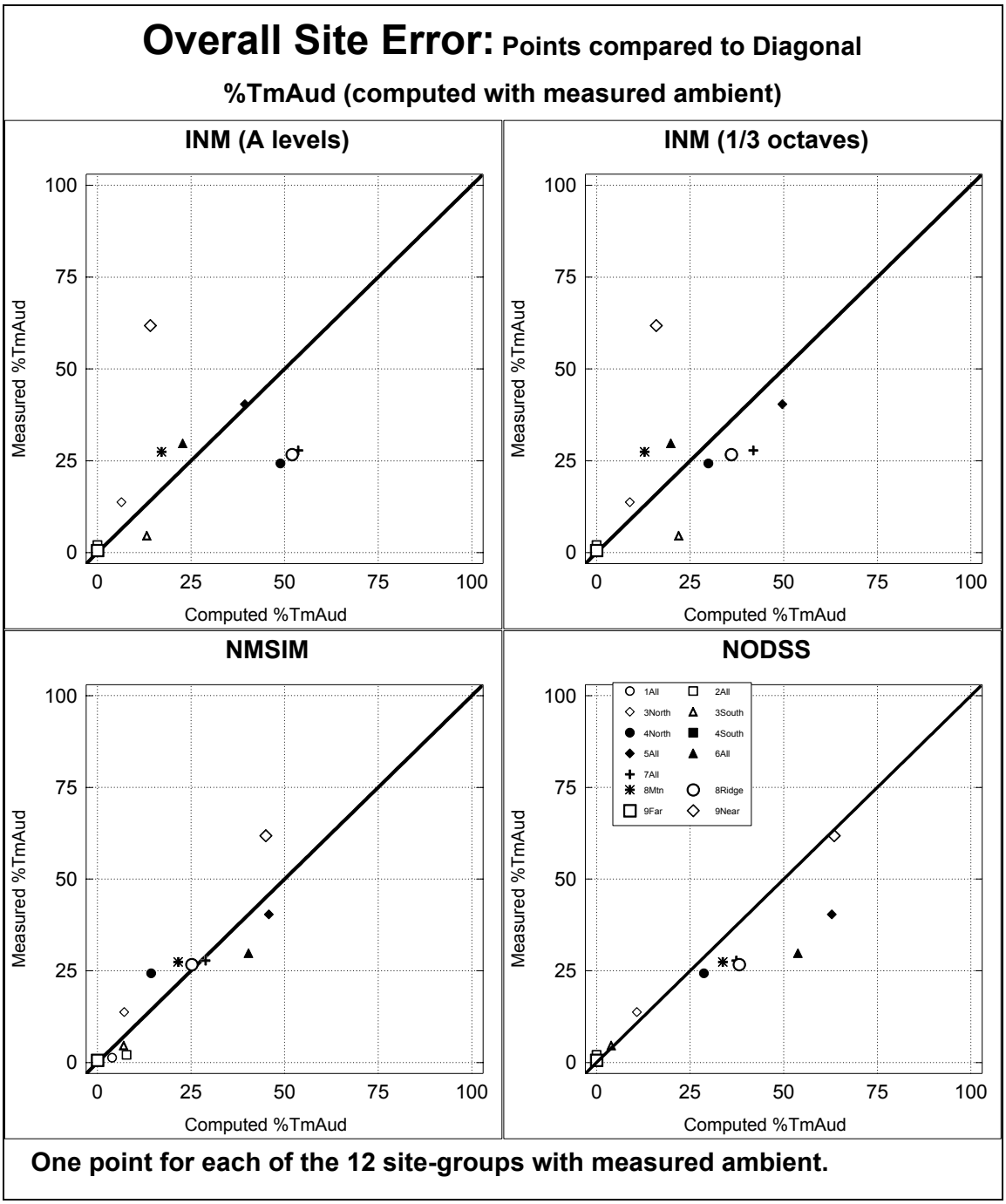


Figure 37. Multi-Hour (Site) Overall Error: Audibility, Computed With Measured Ambient

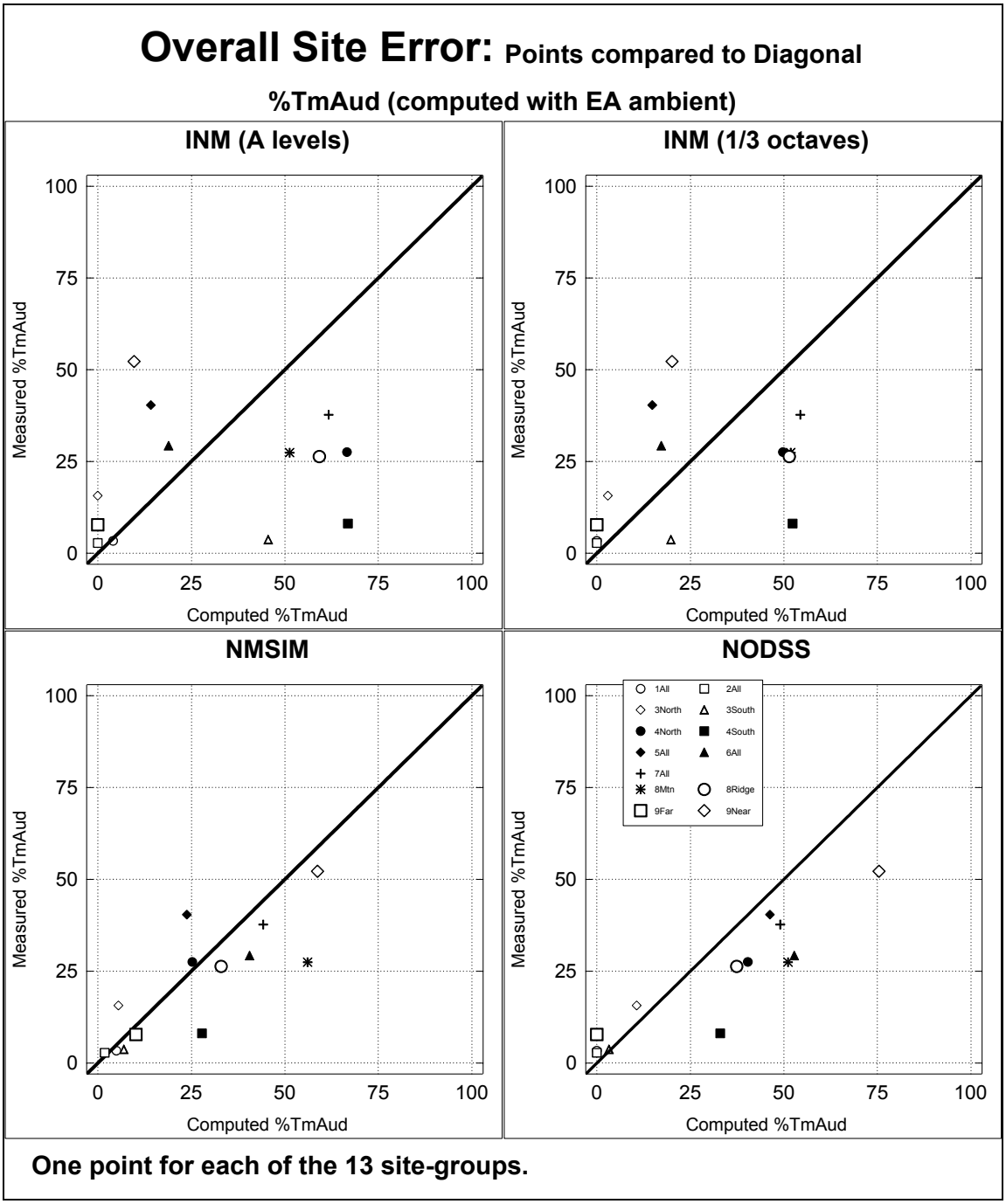


Figure 38. Multi-Hour (Site) Overall Error: Audibility, Computed With EA Ambient

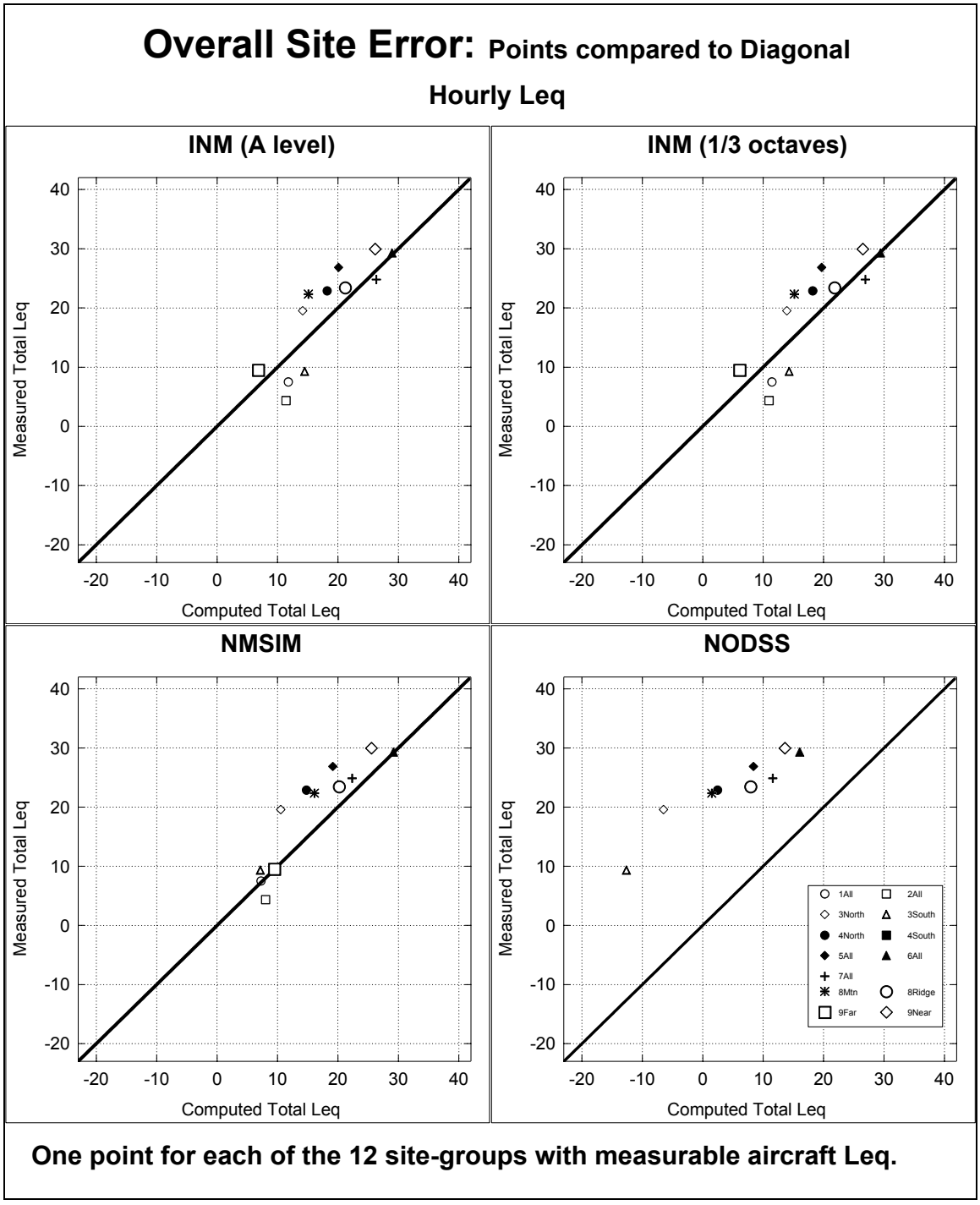


Figure 39. Multi-Hour (Site) Overall Error: Equivalent Level

Table 22. Validation Matrix: Multi-hour (Site-Group) Overall Error

Metric	Ambient sound levels used in computation	Computer model	Overall (rms) error
Audibility	Measured	INM (A levels)	± 16 %TmAud
		INM (1/3 octaves)	± 14 %TmAud
		NMSIM	± 7 %TmAud
		NODSS	± 11 %TmAud
	EA	INM (A levels)	± 30 %TmAud
		INM (1/3 octaves)	± 22 %TmAud
		NMSIM	± 12 %TmAud
		NODSS	± 15 %TmAud
Equivalent Level (L_{eq})	————	INM (A levels)	± 5 dB
		INM (1/3 octaves)	± 5 dB
		NMSIM	± 6 dB
		NODSS	± 19 dB

Comparison of Table 21 and Table 22 confirms that averaging over many hours, to obtain multi-hour comparisons of "measured versus computed," generally reduces overall error.[51] For example, INM(A) single-hour error of 20%TmAud (Table 21, measured ambients) reduces to multi-hour error of 16%TmAud (Table 22)—due to averaging over many hours. Exceptions occur when a model's site-to-site discrepancies are large. For such a model, averaging over many hours does not diminish the overall discrepancy, since it resides in the sites, rather than in the hours. Once again NODSS' bias when computing equivalent levels is clear.

Appendix I.2, page 240, provides additional evidence that error reduces when hourly results are averaged together.

8.4.5 Sources of Measurement Error

8.4.5.1 Approximate adjustment for measurement error due to observer differences

Overall error assesses the overall match between computations and measurements. The previous sections assume that overall error is primarily due to the computation model, rather than to the measurements. However, some measurement error is always present. This section estimates the measurement portion of overall error that could be due to differences in observers and subtracts it out.

Because the two portions of overall error—computation and measurement—are statistically independent of each other, the "variance" of overall error equals the sum of their two variances. Mathematically:

$$\left(\begin{matrix}\text{Overall}\\ \text{error}\end{matrix}\right)^2 = \left(\begin{matrix}\text{Computation}\\ \text{portion}\end{matrix}\right)^2 + \left(\begin{matrix}\text{Measurement}\\ \text{portion}\end{matrix}\right)^2 . \qquad (2)$$

[51] Note that overall error is an "average" value, so it is not automatically smaller for sites (compared to hours) just because fewer points enter the computation.

As a result, the following equation subtracts the measurement portion from overall error.

$$\begin{pmatrix}\text{Computation}\\ \text{portion}\end{pmatrix} = \sqrt{\begin{pmatrix}\text{Overall}\\ \text{error}\end{pmatrix}^2 - \begin{pmatrix}\text{Measurement}\\ \text{portion}\end{pmatrix}^2}. \quad (3)$$

During the study's field measurements, audibility was measured simultaneously at two separate sites—5A and 5B—by two independent teams of listeners who could not see or otherwise observe one another. One person from each team listened at a time, trading off during the day. Figure 40 compares these independent measurements of audibility—one point per listening hour. In the figure, the Volpe-team's measured audibility is plotted horizontally, against the simultaneous measurement of audibility by the NPS team. The apparent measurement errors of both teams are comparable.

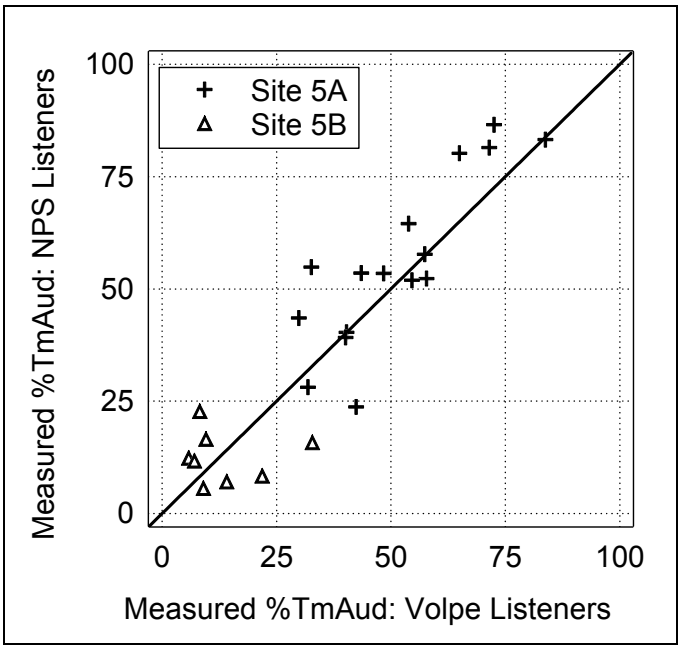


Figure 40. Measurement Portion of Overall Error

From these independent, simultaneous sets of audibility measurements, the rms error is computed to be:

- *Measurement portion of single-hour overall error* approximately equals 11%TmAud.
- *Measurement portion of multi-hour overall error* approximately equals 4%TmAud.

The measurement error of equivalent level was not determined during this study, but instead was estimated:

- *Measurement portion of single-hour overall error* approximately equals 2dB.
- *Measurement portion of multi-hour overall error* approximately equals 1dB.

Since equivalent level was measured with electronic instrumentation, its measurement portion is expected to be relatively small, as estimated. Nevertheless, these estimates are larger than for most

measurements of equivalent level—because the tour-aircraft portion of equivalent level was extracted from the total equivalent level, using observer logs of when tour aircraft dominated.

With Eq.(3) above, Table 23 and Table 24 subtract the measurement portion from the total error, to estimate the portion due to computation, alone.[52]

Table 23. Approximate Computation Portion of Single-hour (Hourly) Overall Error

Metric	Ambient sound levels used in computation	Computer model	Overall (rms) error from Table 21	Approximate measurement portion	Resulting approximate computation portion
Audibility	Measured	INM (A levels)	20 %TmAud	11 %TmAud	17 %TmAud
		INM (1/3 octaves)	19 %TmAud	11 %TmAud	16 %TmAud
		NMSIM	14 %TmAud	11 %TmAud	9 %TmAud
		NODSS	22 %TmAud	11 %TmAud	19 %TmAud
	EA	INM (A levels)	30 %TmAud	11 %TmAud	27 %TmAud
		INM (1/3 octaves)	24 %TmAud	11 %TmAud	21 %TmAud
		NMSIM	17 %TmAud	11 %TmAud	13 %TmAud
		NODSS	20 %TmAud	11 %TmAud	17 %TmAud
Equivalent Level (L_{eq})	————	INM (A levels)	7 dB	2 dB	7 dB
		INM (1/3 octaves)	7 dB	2 dB	7 dB
		NMSIM	8 dB	2 dB	7 dB
		NODSS	18 dB	2 dB	18 dB

Table 24. Approximate Computation Portion of Multi-hour (Site) Overall Error

Metric	Ambient sound levels used in computation	Computer model	Overall (rms) error from Table 22	Approximate measurement portion	Resulting approximate computation portion
Audibility	Measured	INM (A levels)	16 %TmAud	4 %TmAud	16 %TmAud
		INM (1/3 octaves)	14 %TmAud	4 %TmAud	13 %TmAud
		NMSIM	7 %TmAud	4 %TmAud	6 %TmAud
		NODSS	11 %TmAud	4 %TmAud	10 %TmAud
	EA	INM (A levels)	30 %TmAud	4 %TmAud	30 %TmAud
		INM (1/3 octaves)	22 %TmAud	4 %TmAud	22 %TmAud
		NMSIM	12 %TmAud	4 %TmAud	11 %TmAud
		NODSS	15 %TmAud	4 %TmAud	14 %TmAud
Equivalent Level (L_{eq})	————	INM (A levels)	5 dB	1 dB	5 dB
		INM (1/3 octaves)	5 dB	1 dB	5 dB
		NMSIM	6 dB	1 dB	6 dB
		NODSS	19 dB	1 dB	19 dB

The values in Table 23 and Table 24 are approximate and should not be used for model assessment, but are intended to recognize that the errors presented for the various models are likely to be somewhat smaller if measurement error is included.

[52] In both these tables, the right-most column was computed before the other two columns were rounded to the nearest integer. Re-computation of this right-most column, using rounded values as input, may result in values that differ somewhat from those in the table.

8.4.5.2 Possible effects of non-tour aircraft

During measurements, aircraft other than tour aircraft were audible. During training, observers were instructed to log tour aircraft as long as they were audible, regardless of what other sources were audible. Nevertheless, of some 2000 tour events logged across all sites, approximately one-third were either immediately preceded or followed by non-tour aircraft events, of which about 90% were high altitude jets. The presence of these other aircraft (the only non-natural sounds at all sites except site group 6) may have biased some of these tour aircraft observations towards under-measurement of tour audibility if they made the tour aircraft completely inaudible. Two general observations suggest that this bias is likely to be small, or to have no effect on this analysis. First, tour aircraft have audible tonal components to their sound so that broadband jet sound cannot mask tour sound until the jet sound becomes relatively loud. Second, if there is a bias, it affects the analysis of all models equally so that the conclusions of the study about the relative performance of the models would not be changed.

8.4.6 Appreciable site biases

As supplemental insight, Table 25 summarizes the appreciable site biases in the previous three figures. For this summary, appreciable bias is defined as a site discrepancy of more than 10%TmAud, or more than 5dB. These measures of appreciable bias are somewhat arbitrary, but they help identify sites with enough bias that examining the results at these sites in detail might aid in developing model improvements.

Table 25. Appreciable Site Biases

Metric	Ambient sound levels used in computation	Computer model	Appreciable site biases (more than 10%TmAud, or 5dB)	
			Undercomputation	Overcomputation
Audibility	Measured	INM (A levels)	9Near	4North, 7All, 8Ridge
		INM (1/3 octaves)	8Mtn, 9Near	3South, 7All
		NMSIM	9Near	6All
		NODSS		5All, 6All, 8Ridge
Audibility	EA	INM (A levels)	3North, 5All, 9Near	3South, 4North, 4South, 7All, 8Mtn, 8Ridge
		INM (1/3 octaves)	3North, 5All, 9Near	3South, 4North, 4South, 7All, 8Mtn, 8Ridge
		NMSIM		4South, 8Mtn, 8Ridge
		NODSS		4South, 6, 8Mtn, 9Near
Equivalent Level (L_{eq})	————	INM (A levels)	3North, 5All, 8Mtn	9Far
		INM (1/3 octaves)	3North, 5All, 8Mtn	9Far
		NMSIM	3North, 4North, 5All, 8Mtn	
		NODSS	1All, 2All, 3North, 3South, 4North, 4South, 5All, 6All, 7All, 8Mtn, 8Ridge, 9Far, 9Near	

8.5 Model Accuracy: Bias Component of Model Error

8.5.1 Overview

Model accuracy is a measure of a model's performance on the average—that is, how well average computations match average measurements.

If the average of all measured values is subtracted from the average of all computed values, the resulting number—called model bias—quantifies the model's accuracy. This section computes

model biases for each combination of computer model and computed metric. In addition, each model's bias can be diagnosed as a function of various factors. This section examines several such factors. First, bias is diagnosed graphically and numerically as a function of:

- Computed value.

Next, bias is diagnosed graphically as a function of:

- Angle of visibility,
- Vertical temperature gradient,
- Track-to-site wind component, and
- Along-the-track wind component, combined.

These diagnoses are useful to learn whether or not model bias depends upon these factors. When it does, a model may be inappropriate to use for some ranges of the factors. When a model's single-number bias is small and, in addition, it is not biased in any factor range, then we are quite confident of the model's lack of bias and general usefulness to the Park Service.

This section ends with a discussion of possible model calibration and recommendations against it.

8.5.2 Model bias

Model bias is the average of all "computed values minus measured values." Single-hour (hourly) biases result when each site-hour is included explicitly in this averaging, independent of its site-group. Mathematically:

$$\text{Single-hour (hourly) bias} = \frac{1}{H}\left[\sum_{h=1}^{H}\left(S_{h,\,\text{computed}} - S_{h,\,\text{measured}}\right)\right]. \tag{4}$$

In this equation, S_h is the sound metric—computed and measured. The averaging is over all H site-hours in the study, independent of site-group.

Multi-hour (site) biases result when values are first averaged within each site-group, and then the site-group results are averaged together. Mathematically:

$$\begin{aligned} B_g &= \text{each site-group's bias} \\ &= \frac{1}{H_g}\left[\sum_{h=1}^{H_g}\left(S_{h,\,\text{computed}} - S_{h,\,\text{measured}}\right)\right], \text{ and then:} \\ \text{Multi-hour (site) bias} &= \frac{1}{G}\left[\sum_{g=1}^{G}\left(B_g\right)\right]. \end{aligned} \tag{5}$$

In this equation, S_h is again the sound metric—computed and measured. The first averaging is over all H_g site-hours in that site-group. Then all the site-group biases are averaged in the second part of this equation, in which G is the number of site-groups.

Resulting single-number biases of both types—single hour and multi-hour—appear in Table 26. Each bias in this table is accompanied by its 95-percent confidence limits. Large data scatter produces large confidence limits.[53]

Table 26. Validation Matrix: Model Biases

Metric	Ambient sound levels used in computation	Computer model	Bias ±95-percent confidence range	
			Single-hour (hourly)	Multi-hour (site)
			Value	Value
Audibility	Measured	INM (A levels)	+3 ± 10	+1 ± 12
		INM (1/3 octaves)	+1 ± 8	–2 ± 10
		NMSIM	+1 ± 4	–1 ± 4
		NODSS	+10 ± 6	+6 ± 5
	EA	INM (A levels)	+1 ± 17	+5 ± 17
		INM (1/3 octaves)	–2 ± 13	+1 ± 13
		NMSIM	–1 ± 7	+2 ± 6
		NODSS	+10 ± 5	+8 ± 6
Equivalent Level (L_{eq})	———	INM (A levels)	–2 ± 2	–1 ± 3
		INM (1/3 octaves)	–2 ± 3	–1 ± 3
		NMSIM	–4 ± 2	–3 ± 2
		NODSS	–18 ± 3	–26 ± 8

8.5.3 Diagnosis of bias

8.5.3.1 Diagnosis by computed value

This section diagnoses model bias by each hour's computed value. This diagnosis starts with a graphical comparison of "measured versus computed," supplemented by regression lines through these plotted points. From these regression lines and their 95-percent confidence limits, it then computes each model's bias, which varies for different computed values.

Figure 41 through Figure 43 show plots of "measured versus computed" data, with a point for each measured site-hour. These figures duplicate three of the previous figures, but add regression lines and their 95-percent confidence limits.[54] The lighter curved lines are the regression lines, while the

[53] Note that single-hour and multi-hour biases in this table often have nearly the same values and nearly the same confidence limits. In fact, if all site-groups had the same number of hours, these values would be exactly the same—for the following reason. These two types of single-number biases are just slightly different ways of averaging "computed minus measured." In particular, the single-hour biases in the table are pure hourly averages, completely independent of site-groups. Each *hour* is counted (weighted) equally. In contrast, the multi-hour biases count (weight) each *site-group* equally—even those with very few hours. That is the reason for the differences in these two types of single-number bias—different numbers of hours in each site-group.

[54] For audibility, these regression lines and their confidence limits were computed with the computer program MLwiN—Jon Rasbash et al, *A User's Guide to MLwiN,* Multilevel Models Project, Institute of Education, University of London, 2000. This computer program is based upon the multilevel statistics in Harvey Goldstein, *Multilevel Statistical Models, Second Edition*, Kendall's Library of Statistics 3, Oxford University Press, New York, 1995. Alternative programs exist for comparable computations—in medical and sociological research, which both involve complex data sampling.

heavier bounding lines are their confidence limits. They are heavier here because they are more important to the computation of model accuracy.

The regression lines in these figures show the average relationship between the site-hour computations (horizontal axis) and their corresponding measurements (vertical axis). As each regression line progresses from left to right, it passes approximately through the center of the vertical point scatter—that is, through the vertical average. For this reason, these lines graphically show how well computations match measurements, on the average. The diagonal line indicates a perfect match.

We are 95-percent confident that the actual vertical average (all possible hours at all possible sites) lies somewhere between the two heavy lines that bound the regression. As these bounding lines progress from left to right, they sometimes encompass the diagonal, and sometimes do not. Where they do, the match between computations and measurements is "accurate," with 95-percent confidence. Where the diagonal is outside these bounding lines, model computations are "biased" — that is, the average computation does not match average measurements with 95-percent confidence.

Table 27 contains the results of this regression analysis. As discussed above in Section 8.2.2, accuracy was assessed at four locations along the horizontal axes of Figure 41 through Figure 43. These computed values appear in the table's fourth column. The fifth column then contains the model's bias range at each of these computed values. These bias ranges show the relation between each plot's diagonal line (perfect agreement) and its 95-percent confidence bounds on the regression:

MLwiN properly accounts for the study's sampling method (sampled measurement sites, then sampled measurement hours at those sites). As a result, its computed confidence limits take into account both the hour-to-hour variability and the site-to-site variability in "measured versus computed." These confidence bands widen somewhat when hour-to-hour variability is large—but not extremely, because hourly variability averages out over many measured hours. In contrast, they widen extremely when site-to-site variability is large—because site variability averages out over only 13 sites. MLwiN takes both variabilities into account, including their relative importance.

Another way of understanding multilevel analysis hinges on the concept of "independent data points." Confidence limits computed with MLwiN properly account for the number of truly independent data points in the regression. For a computer model with absolutely no site bias, for example, all the data points in the plots are truly independent. However, for a computer model with very large site biases, each site's hourly data points are highly correlated with each other, so that we truly only have 13 independent comparisons of computations with measurements, one comparison per site. Multi-level regression sorts this out, depending upon the within-site and between-site correlation it finds in the data points.

Within MLwiN, we chose to use logistic regression, instead of linear regression, to obtain these regression lines, for the following reason. Exploratory linear regression produced 95-percent confidence bounds that sometimes went below zero percent and/or above 100 percent. Such impossible results are clear evidence that linear regression is improper here. To produce such impossible results, linear regression's "Gaussian" assumption about the underlying vertical scatter has to be substantially false. In contrast, logistic regression assumes all data points lie between zero and unity (0 and 100 percent), and so it produces confidence bands also limited between 0 and 100 percent. Its "binomial" assumption about the underlying vertical scatter is far closer to truth. With this assumption, MLwiN obtains regression results with maximum-likelihood mathematical methods.

For equivalent level, these regression lines and their confidence limits were computed with the computer program Statistica—*Statistica for Windows (Computer Program Manual),* StatSoft Inc., www.statsoft.com, Tulsa OK, 1999. This computer program produces linear regressions and their confidence limits in the normal manner (least-squares method), assuming Gaussian scatter.

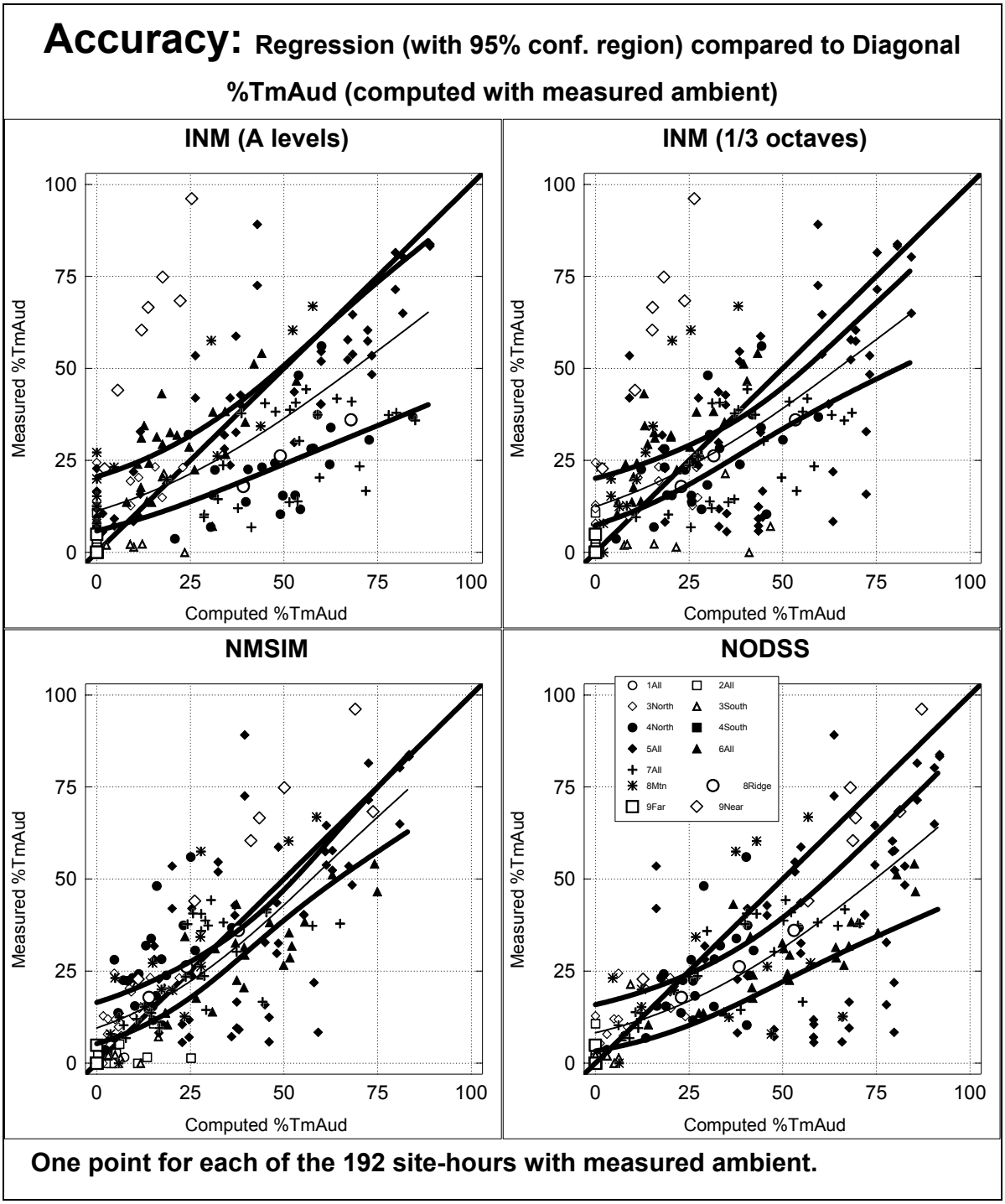


Figure 41. Bias Diagnoses With Computed Value: %TmAud, Computed With Measured Ambient

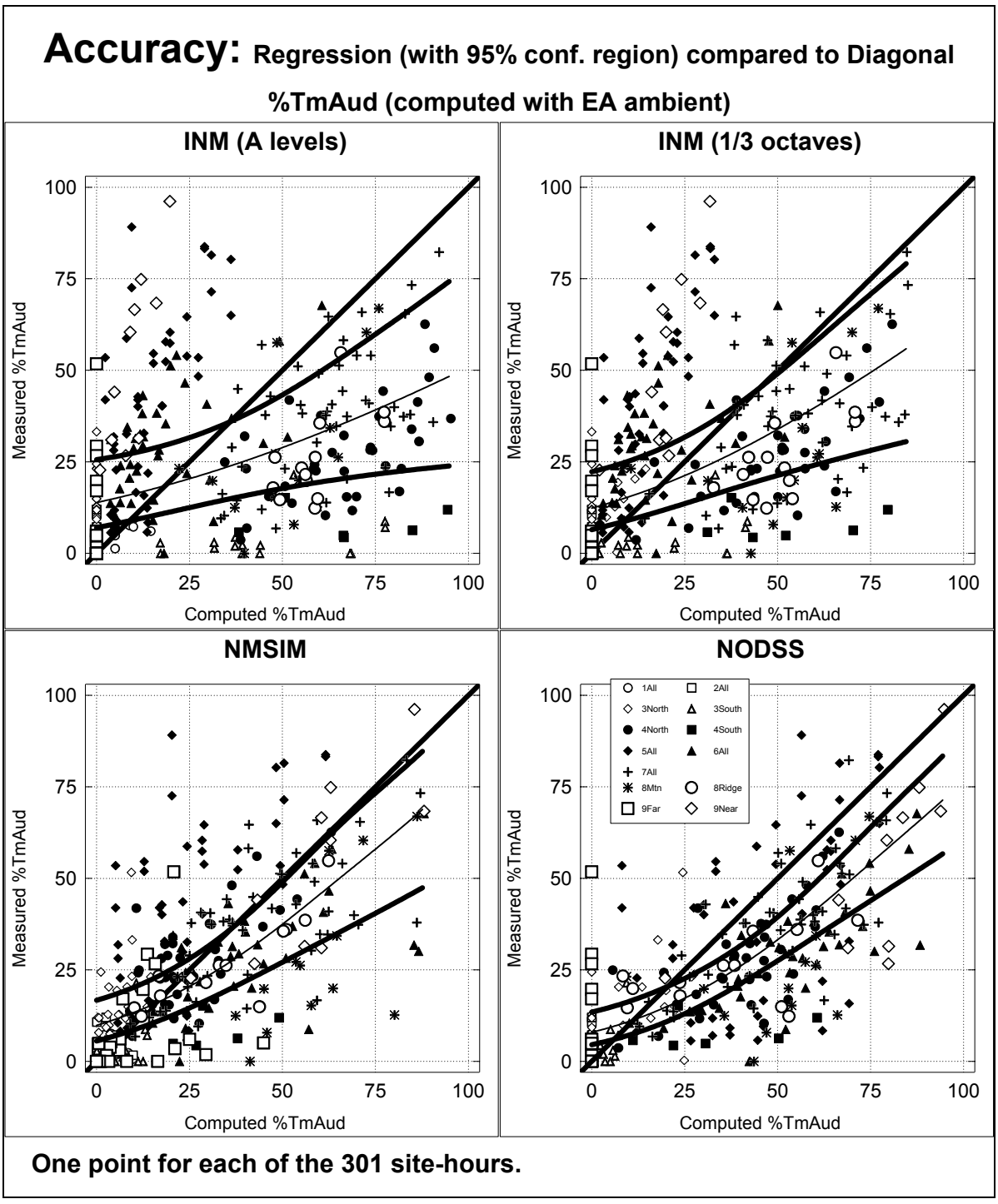


Figure 42. Bias Diagnoses With Computed Value: %TmAud, Computed With EA Ambient

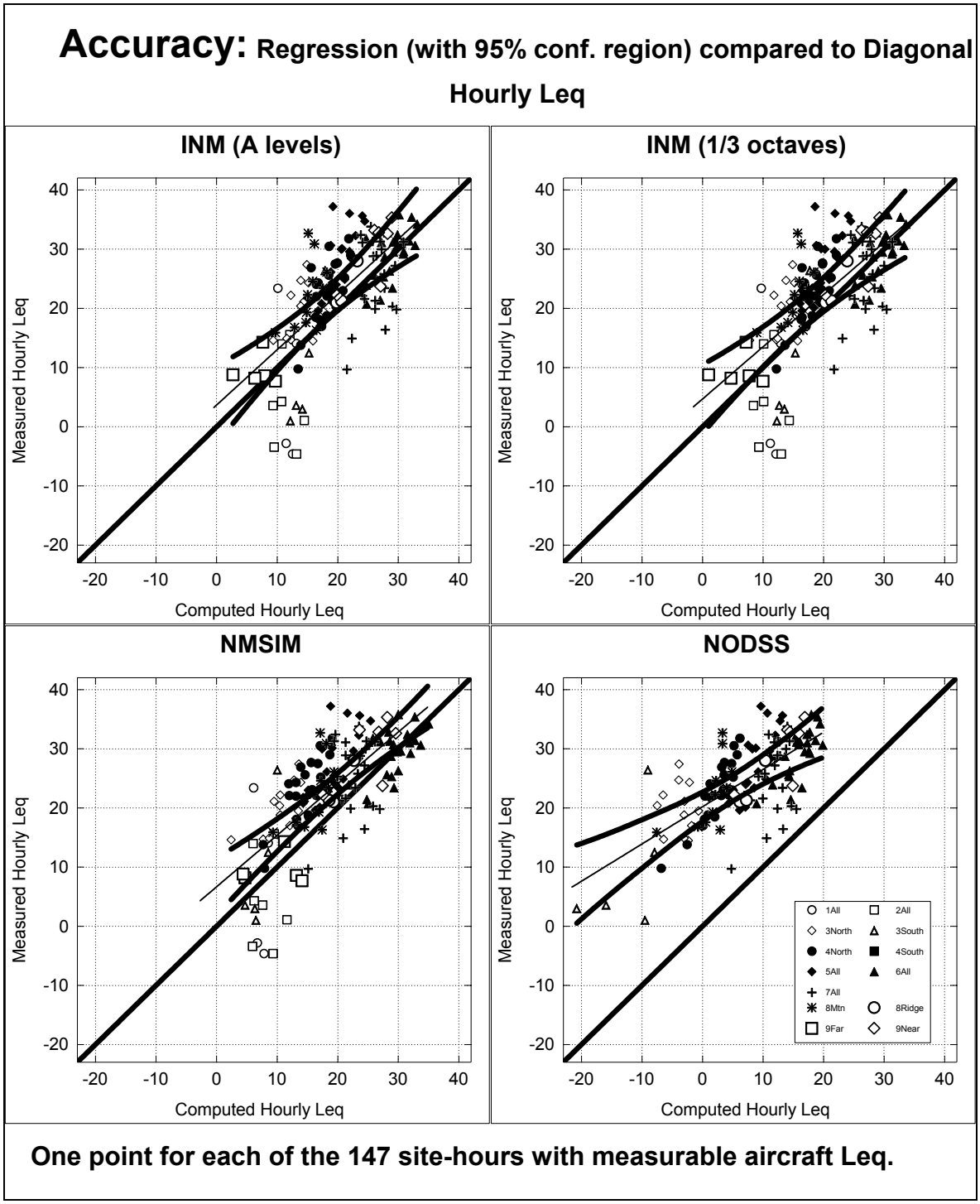


Figure 43. Bias Diagnoses With Computed Value: Equivalent Level (L_{eq})

- Where the diagonal line is not within the confidence bounds, both algebraic signs of the bias range are the same, and therefore we are 95-percent confident that the model's bias lies somewhere within this range.
- In contrast, where the diagonal does lie within the confidence bounds, we are not 95-percent confidant that the model is truly biased (for this computed value). Instead, the apparent bias (the diagonal-to-regression vertical offset) may be the result of random sampling error.

Table 27. Validation Matrix: Bias Diagnosis with Computed Value

Metric	Ambient sound levels used in computation	Computer model	Model accuracy	
			Computed value	Bias range: computed minus measured values, on the average
			%TmAud	%TmAud
Audibility	Measured	INM (A levels)	0	–20 to –10
			25	–6 to +12
			50	0 to +26
			75	+2 to +40
		INM (1/3 octaves)	0	–20 to –8
			25	–4 to +6
			50	+4 to +16
			75	+6 to +28
		NMSIM	0	–16 to –6
			25	–4 to +8
			50	+2 to +12
			75	0 to +18
		NODSS	0	–16 to –4
			25	0 to +14
			50	+10 to +28
			75	+12 to +40
	EA	INM (A levels)	0	–26 to –8
			25	–6 to +12
			50	+6 to +32
			75	+14 to +54
		INM (1/3 octaves)	0	–22 to –6
			25	–6 to +12
			50	0 to +28
			75	+4 to +46
		NMSIM	0	–16 to –6
			25	–4 to +10
			50	0 to +22
			75	+2 to +34
		NODSS	0	–14 to –4
			25	+2 to +12
			50	+10 to +22
			75	+12 to +32
			dB	dB
Equivalent Level (L_{eq})	————	INM (A levels)	5	–6 to +2
			20	–6 to 0
			30	–6 to +2
		INM (1/3 octaves)	5	–8 to 0
			20	–6 to 0
			30	–6 to +4
		NMSIM	5	–8 to –2
			20	–6 to –2
			30	–6 to 0
		NODSS	–5	–26 to –20
			5	–20 to –16
			15	–18 to –12

8.5.3.2 Diagnoses by physical factors

This section diagnoses model bias by various physical factors in the study. These diagnoses also start graphically, in plots of "computed minus measured" versus each of these other factors. Because "computed minus measured" is each point's contribution to bias, these plots essentially graph bias vertically against these other factors horizontally.

These diagnostic plots diagnose bias by the following physical factors:

- *Angle of visibility:* Figure 44 through Figure 46, separately for each metric,
- *Vertical temperature gradient:* Figure 47 through Figure 49,
- *Track-to-site wind component:* Figure 50 through Figure 52, and
- *Along-track wind component:* Figure 53 through Figure 55.

In each figure, points are distinguished by their site distances from the flight track, as shown in the key. In each panel of these figures, locally weighted regression lines are shown to help visualize the central tendency of the plotted points, separately by distances from the flight track.

These figures lead to the following conclusions concerning model bias:

- *Angle of visibility.* In the INM panels for audibility (upper panels of Figure 44 and Figure 45), the regression lines rise upward to the left. This regression-line pattern indicates that INM overcomputes audibility (upward) when both (1) visible angle is small (leftward in the plots) and also (2) distance is small (the regression line for solid circles and open triangles). Therefore,

  > *INM (both versions) appears to overcompute audibility close to the flight track, for sites that are shielded by terrain from the flight track.*

  This point and regression-line pattern is somewhat less visible in the INM panels of Figure 46, though there appears still to be the tendency for overcomputation at smaller angles of visible corridor and undercomputation at larger angles. This regression-line pattern (upward to the left) does not exist in the bottom panels of these figures.

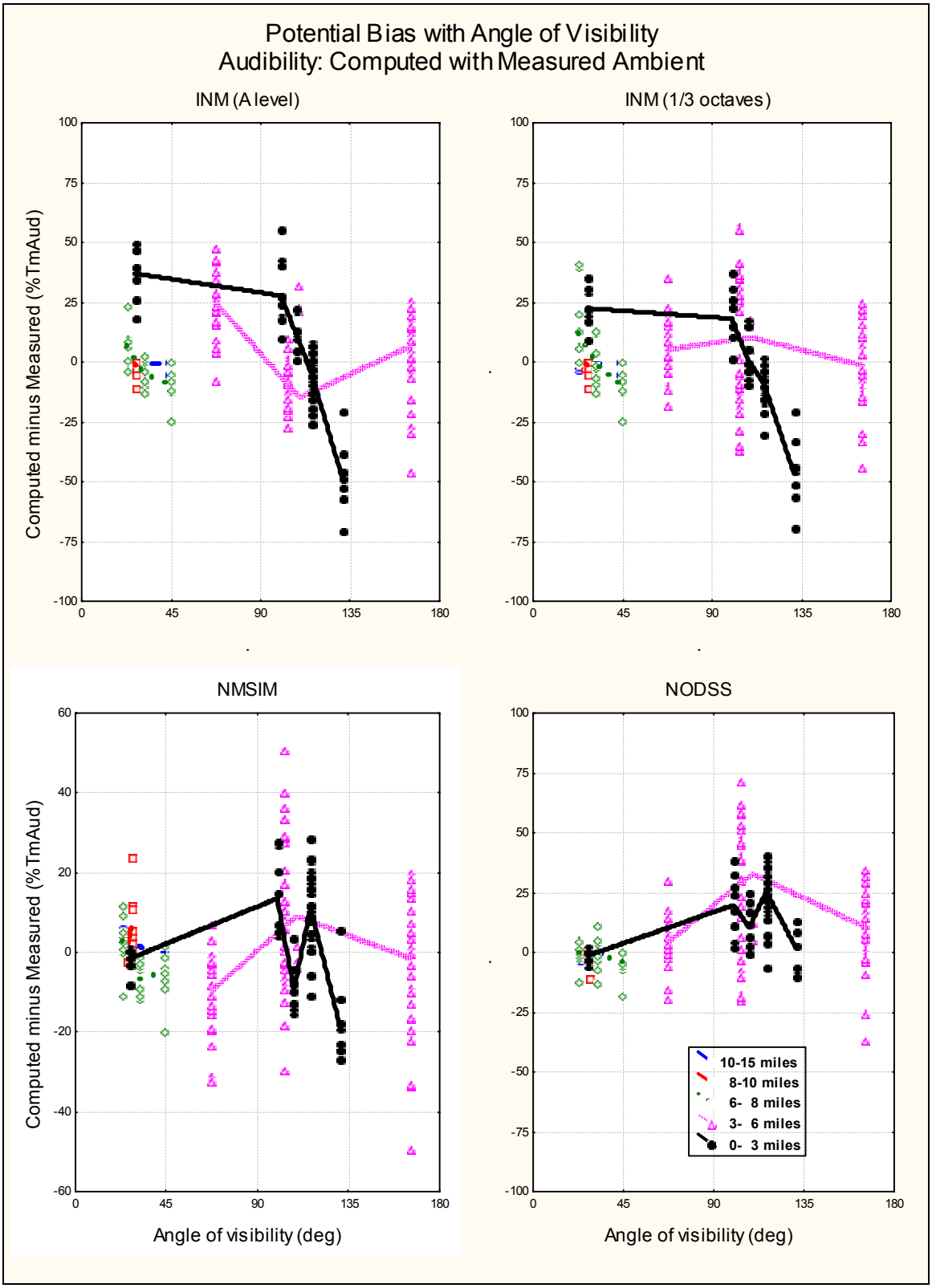


Figure 44. Effect of Angle of Visibility on Model Bias: Audibility, Measured Ambient

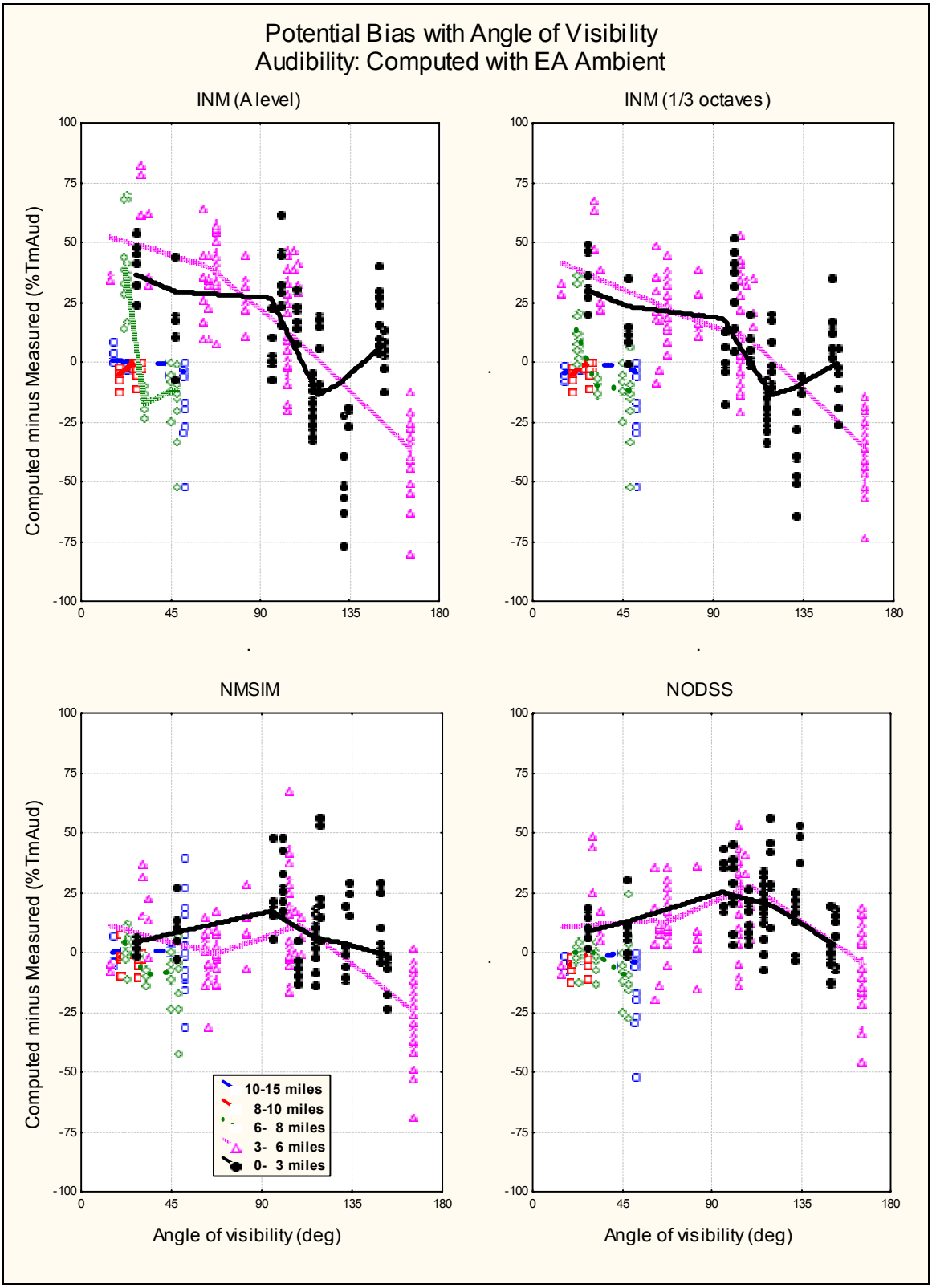


Figure 45. Effect of Angle of Visibility on Model Bias: Audibility, EA Ambient

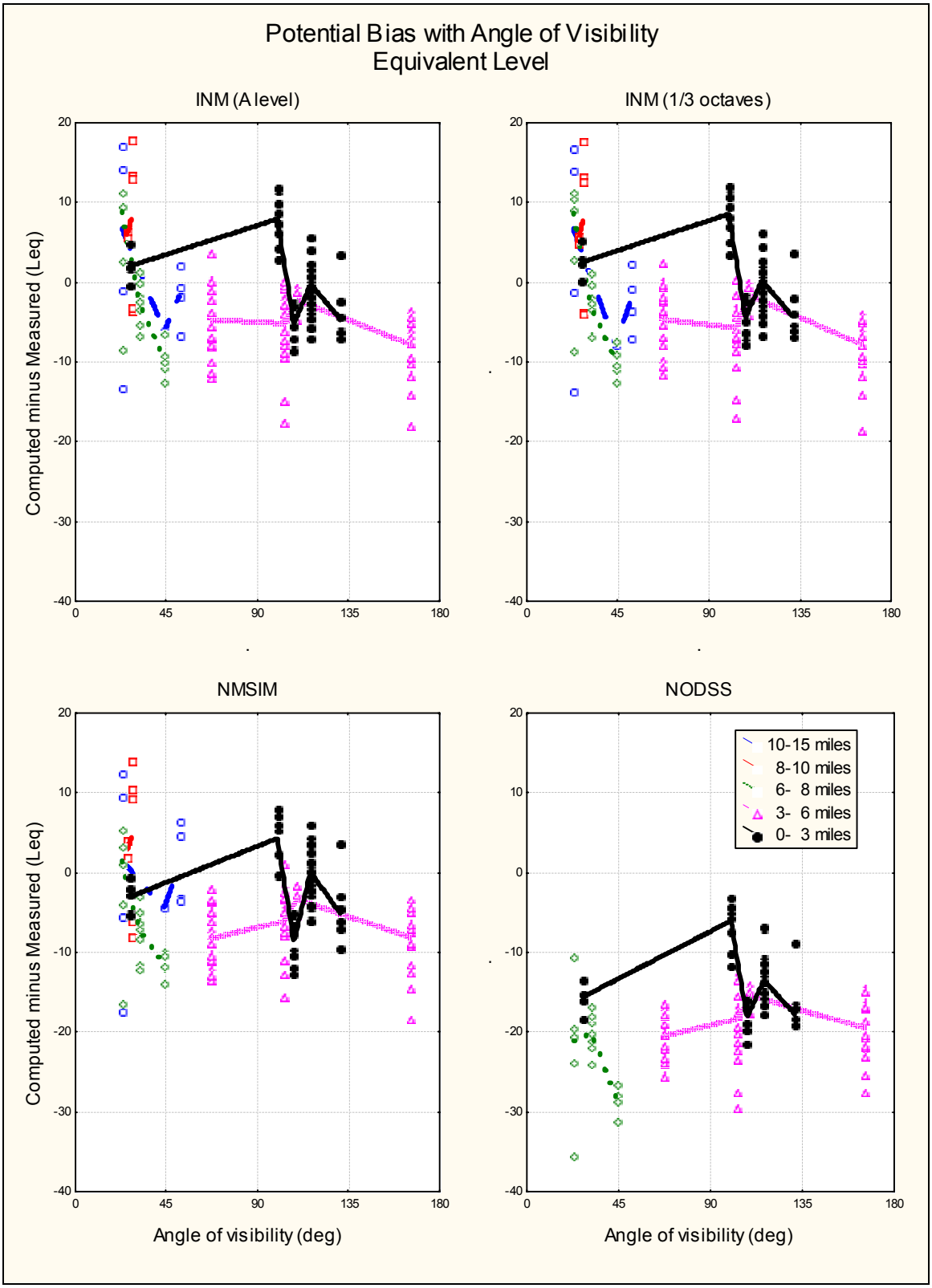


Figure 46. Effect of Angle of Visibility on Model Bias: Equivalent Level

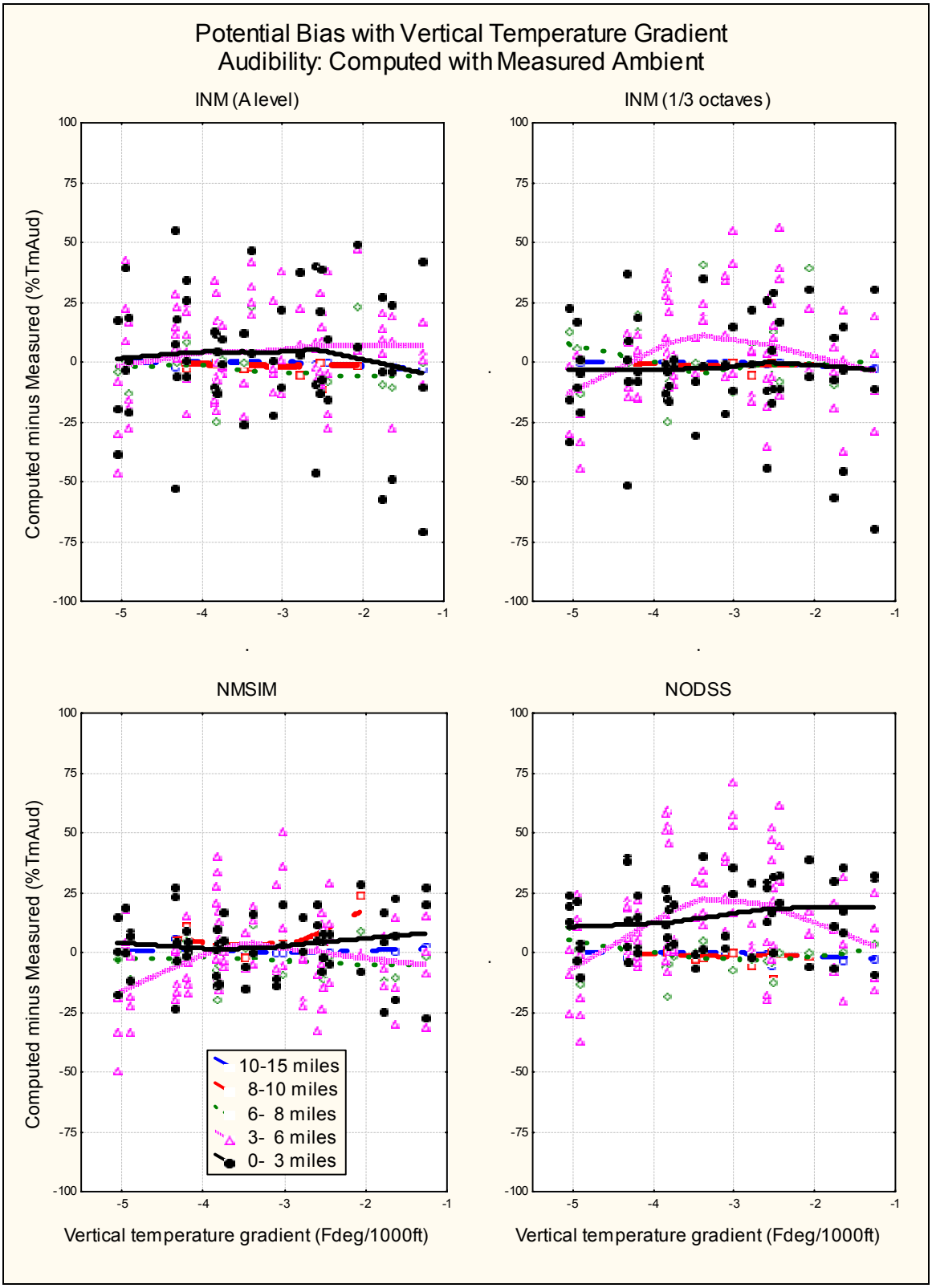


Figure 47. Effect of Vertical Temperature Gradient on Model Bias: Audibility, Measured Ambient

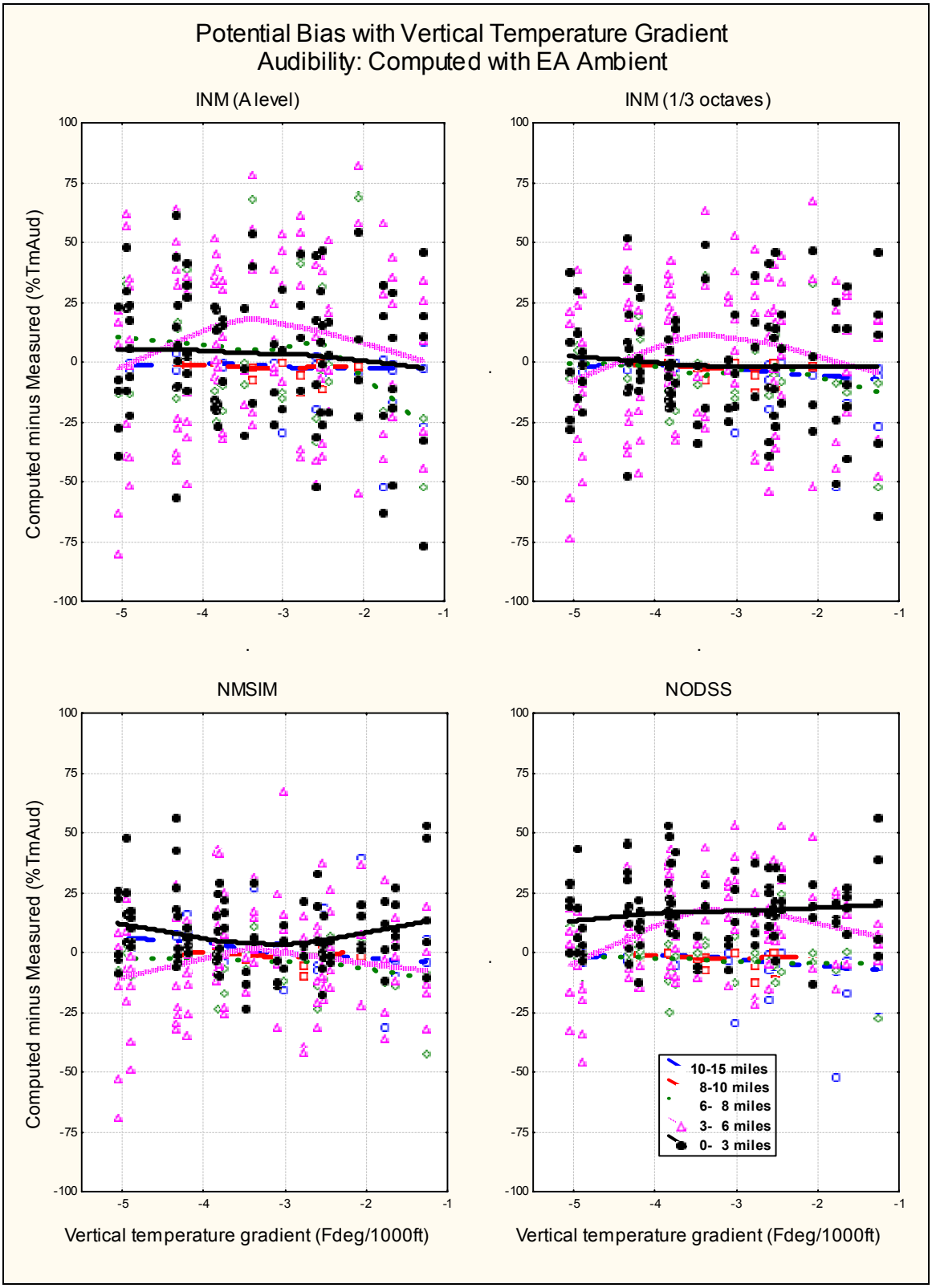


Figure 48. Effect of Vertical Temperature Gradient on Model Bias: Audibility, EA Ambient

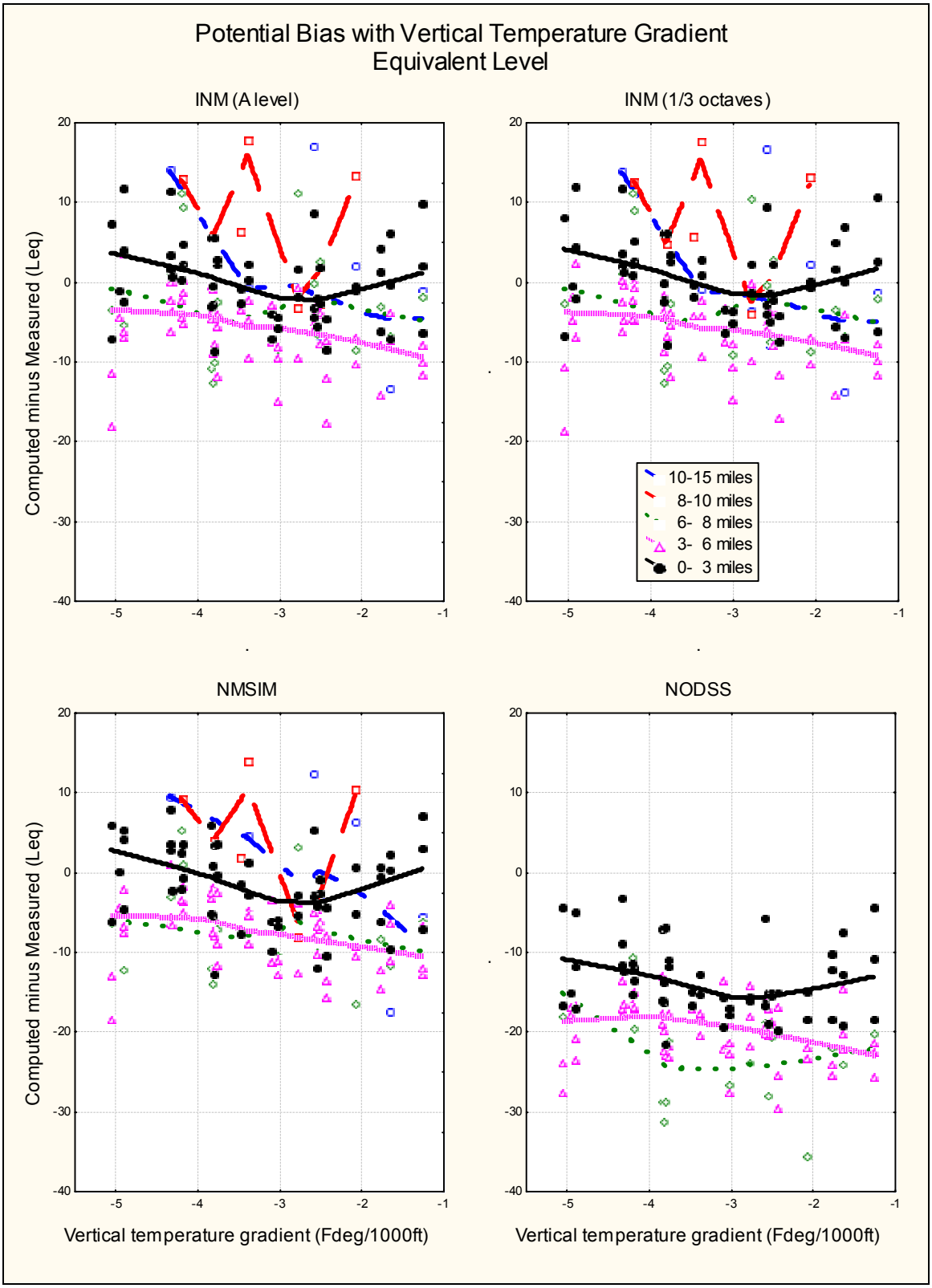


Figure 49. Effect of Vertical Temperature Gradient on Model Bias: Equivalent Level

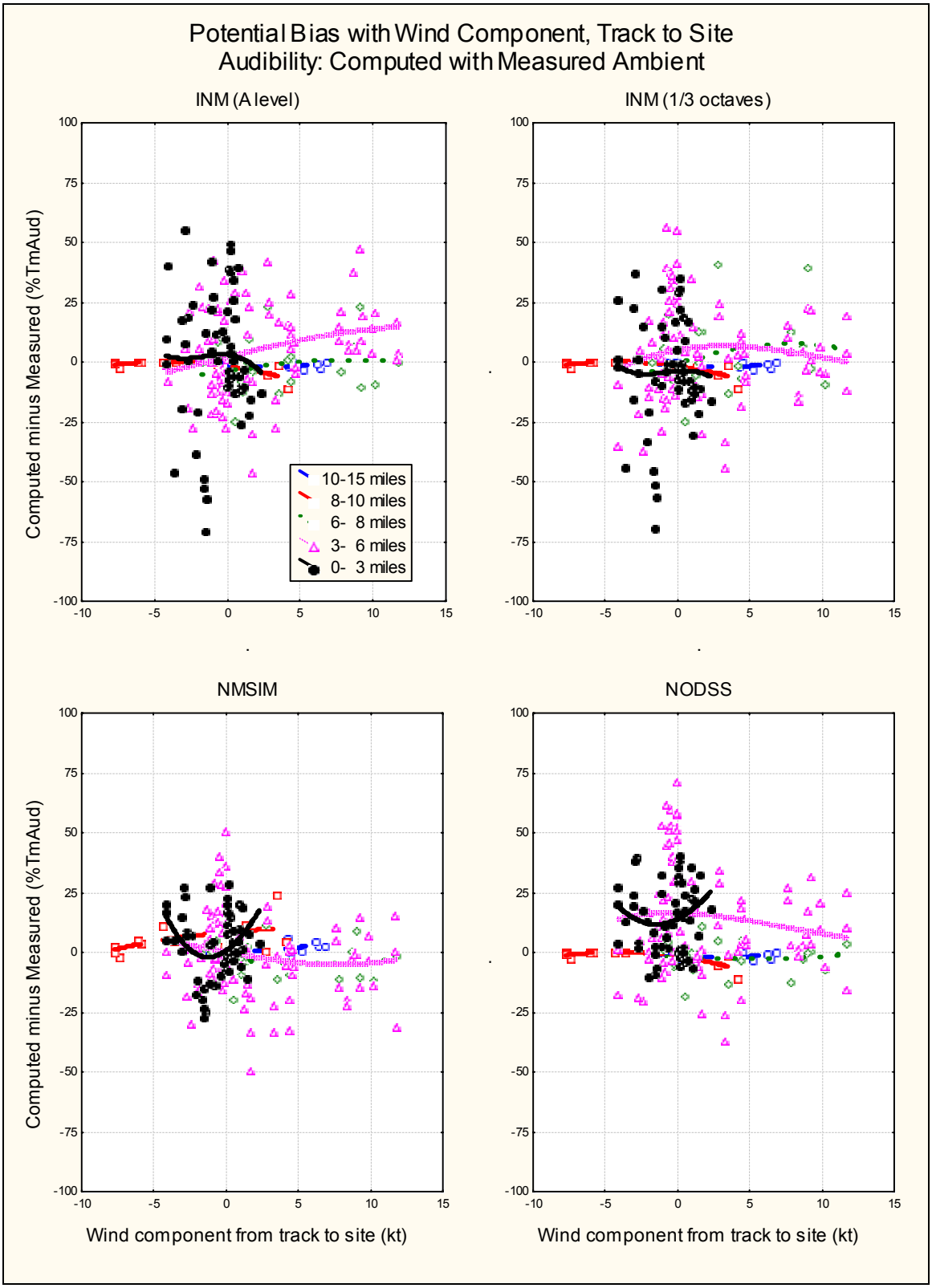


Figure 50. Effect of Track-to-Site Wind Component on Model Bias: Audibility, Measured Ambient

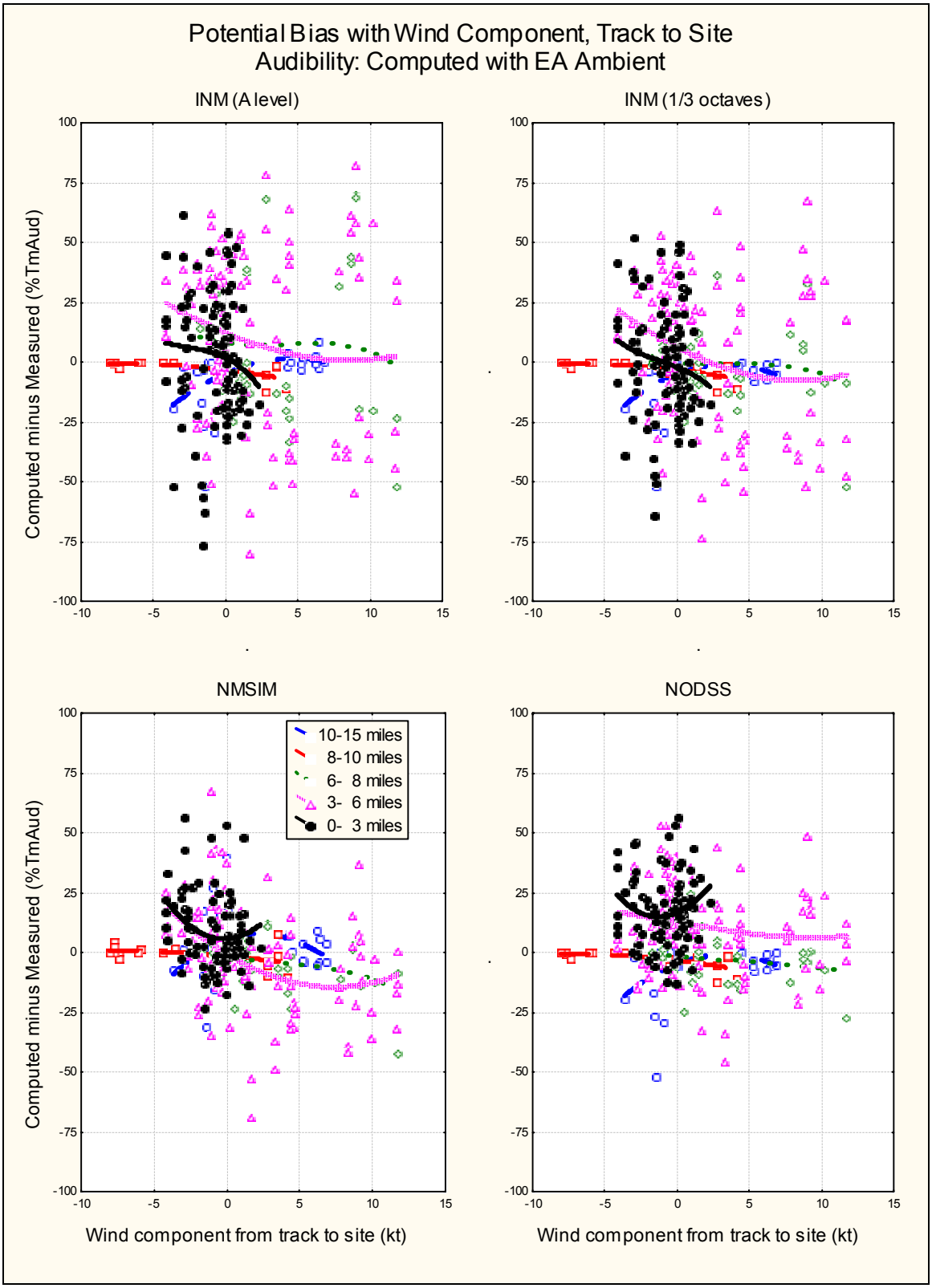


Figure 51. Effect of Track-to-Site Wind Component on Model Bias: Audibility EA Ambient

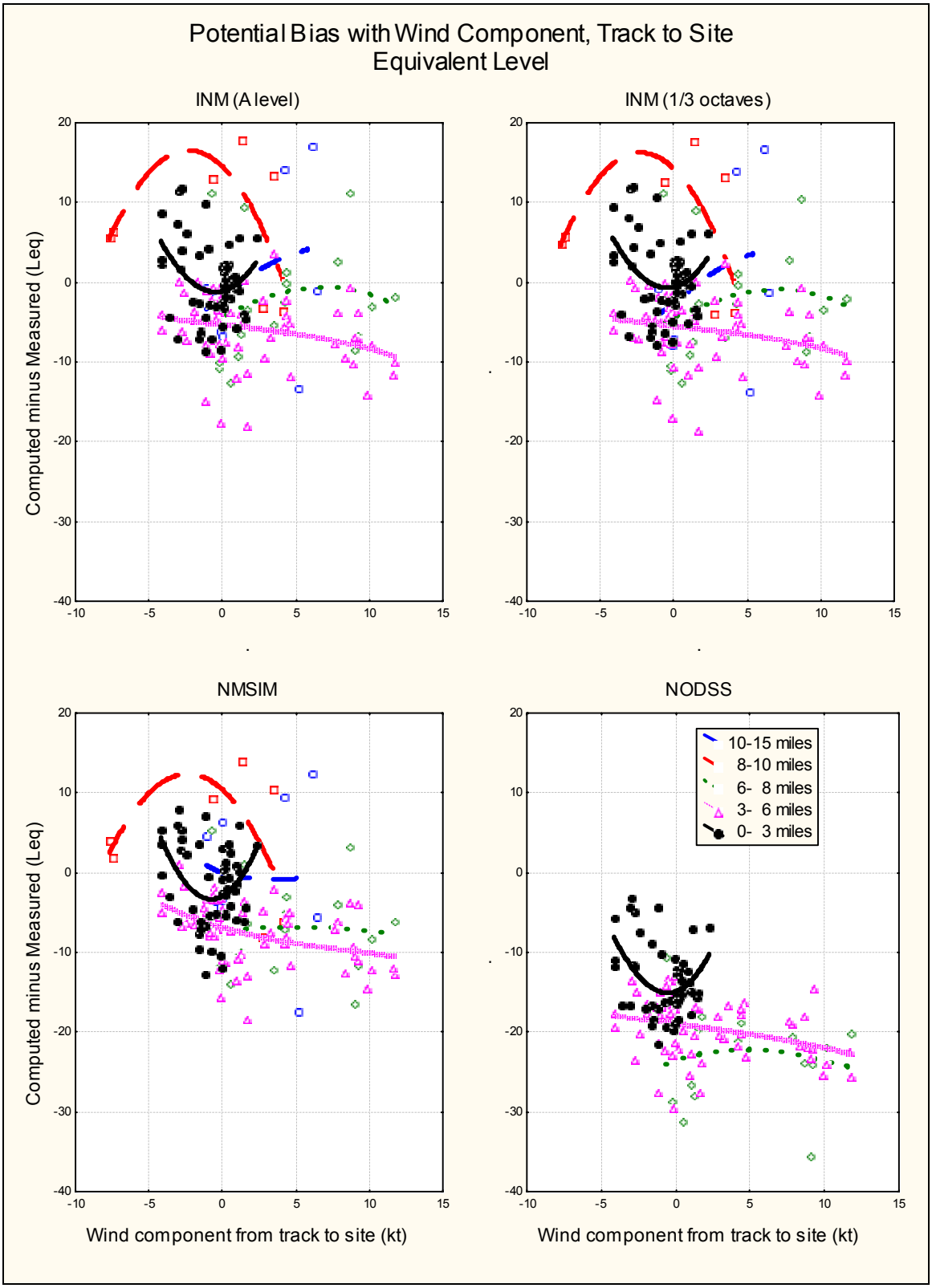


Figure 52. Effect of Track-to-Site Wind Component on Model Bias: Equivalent Level

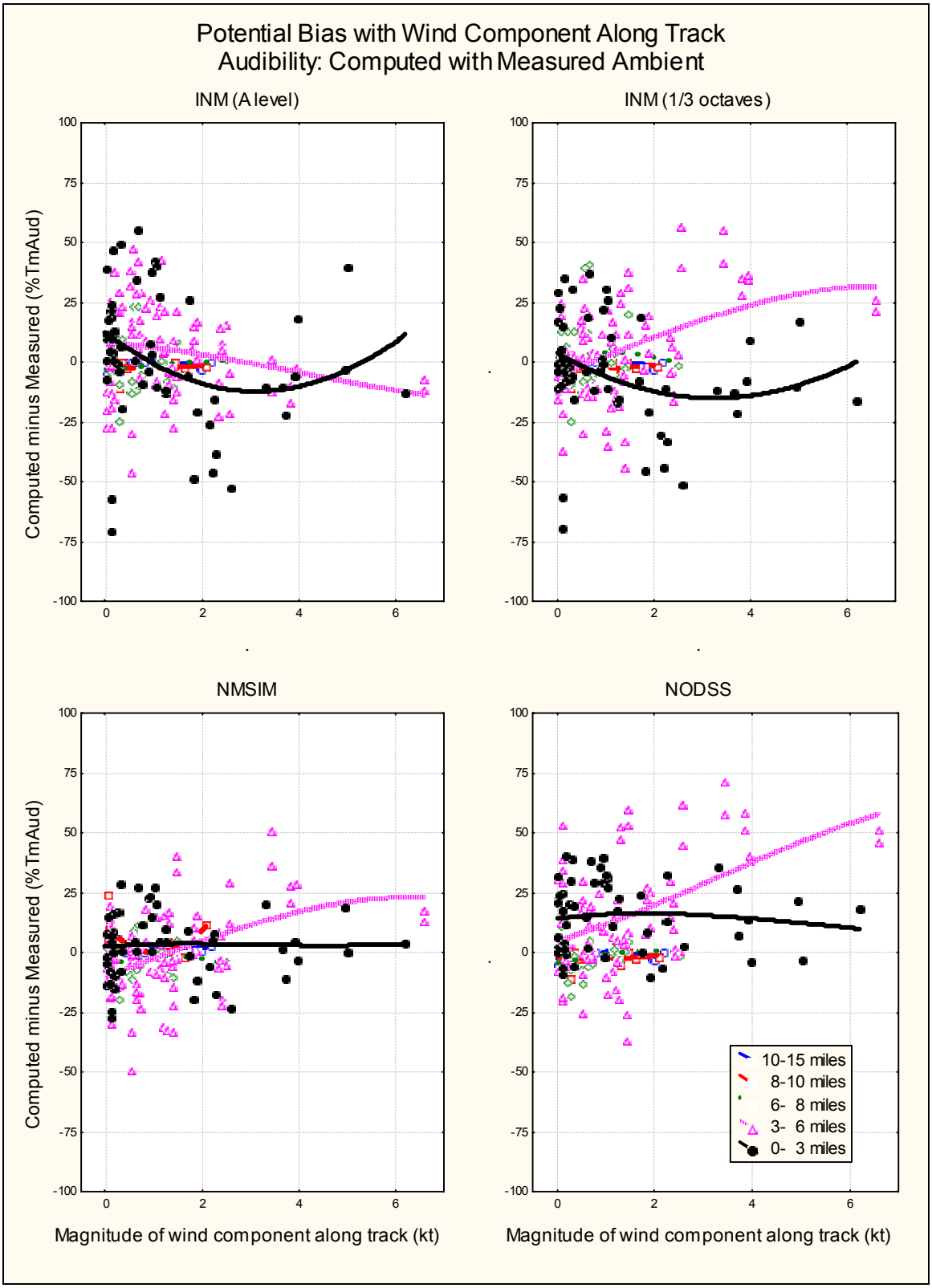


Figure 53. Effect of Along-Track Wind Component on Model Bias: Audibility, Measured Ambient

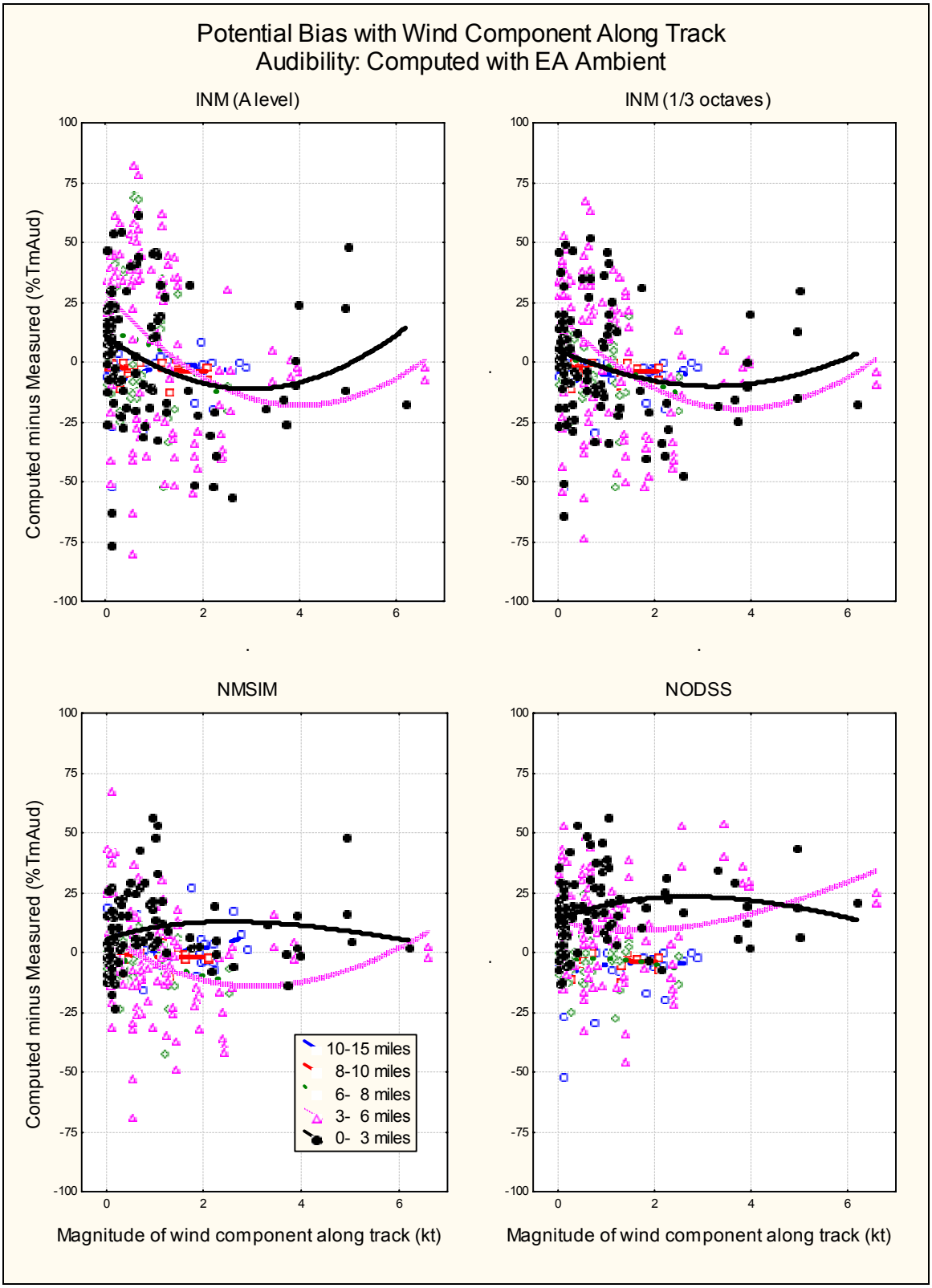


Figure 54. Effect of Along-Track Wind Component on Model Bias: Audibility, EA Ambient

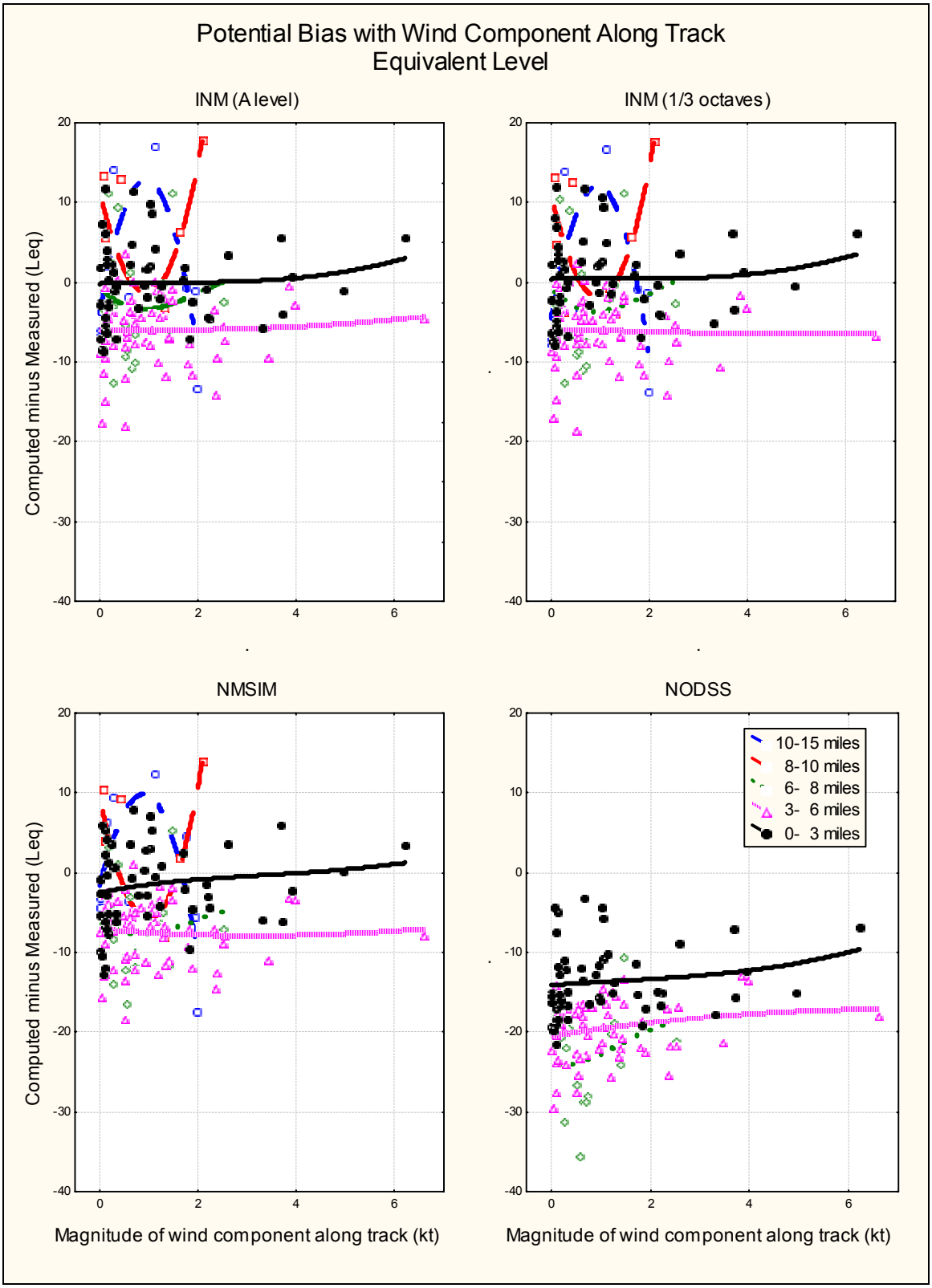


Figure 55. Effect of Along-Track Wind Component on Model Bias: Equivalent Level

- *Vertical temperature gradient.* All panels of Figure 49 (equivalent level) show points and regression lines that generally rise slightly upward to the left—though less for NODSS than for the other models. Therefore,

  > *all models appear to overcompute equivalent level when vertical temperature gradients are most negative (further to the left in the plots).*[55]

  This trend is most apparent for sites at larger distances (open circles and open squares). This same upward-to-the-left pattern is less apparent for audibility (Figure 47 and Figure 48).

- *Track-to-site wind component.* None of the points or regression lines in Figure 50 through Figure 52 shows clear upward or downward trends. Therefore, model bias does not seem to depend upon this wind component.

- *Along-track wind component.* Some of the models in Figure 53 through Figure 55 show slight upward-to-the right trends for some of the points and regression lines, but not for all. As a result, major systematic bias is not apparent from these plots.

8.5.4 Possible model calibration and recommendation against it

Model calibration means adjusting a model's output so that it computes well on the average—that is, without bias. Because the accuracy regressions determine bias, they also determine the appropriate numerical calibration to eliminate this bias.

Calibration was originally a part of this study's goal. Calibration, as discussed in Section 1.9.1.2 with respect to Figure 1, is the forced removal of bias in a model. However, due (1) in part to some of the models providing what is judged to be reasonable levels of accuracy and precision, but (2) due mainly to the shortcomings of resorting to this type of calibration, calibration is not recommended. This type of calibration must rely solely on the data used and on the model to be calibrated, and takes no account of possible reasons for discrepancies. Hence, a calibrated model provides little certainty that its use for different conditions or for different parks will provide realistic results.[56] It is recommended that rather than resorting to calibration, models be used as they currently are configured, or that improvements be made to the models as appropriate. (Section 1.11.2 or Section 11.2 summarizes the areas of the models recommended for examination and possible improvement.)

8.6 Model Precision: Random Component of Model Error

8.6.1 Overview

As discussed in the previous section, model *accuracy* measures a model's *average* performance—that is, how well model computations match measurements for each measured hour, or for each

[55] This is what would be expected from vertical temperature gradients during daytime. Temperature lapse (temperature decreasing with altitude) produces upward refraction, which reduces sound levels at larger distances—especially when the source of sound is in direct view. This sound-level reduction is not in the models, so they overcompute. The effect is expected to be less when terrain intervenes.

[56] Calibration is often acceptable when it is based on physical reasons. For example, the appropriate value for one of the variables in a model may be unknown, such as sound attenuation due to forests. If measurements are taken in such a way to yield a valid comparison of forest and non-forest attenuation, then the results might be used to quantify the forest attenuation and hence "calibrate" it for forests. Both the INM and NODSS as applied in this study, use a type of calibration. Neither model internally accounts for overlapping sound of closely spaced aircraft; the audibility time for each aircraft is computed independently. To account for this possible over-prediction of audibility, an empirical adjustment was applied to INM and NODSS results, see APPENDIX J page243.

measured site-group. Averages over many hours and many site-groups are not always appropriate, however. Sometimes computations are needed for an individual hour, or for an individual site. Whenever hourly or site-group averages are not appropriate, then model *precision* is important.

Model *precision* concerns model performance for each individual hour, or each individual site-group, compared to the average. A model is more precise if it closely matches every single measurement, rather than just the average—that is, if it has very small scatter about the average. This section assesses model precision.

8.6.2 Mathematical computation

Mathematically, model precision has two numeric measures:

- The standard deviation of the scatter (residuals) around the regression line in Section 8.5.3.1, above.
- The correlation coefficient between measured and computed values.

Both of these measures were computed in the standard manner by the computer program *Excel*. For single-hour (hourly) precision, all site-hours are included in this computation. For multi-hour (site) precision, only site-group average points are included, one per site-group.

8.6.3 Resulting precision and random error

Table 28 contains the resulting model precisions, separately by the type of sound metric, the ambient sound levels used in computation, and by computer model. Single-hour (hourly) values are relevant when a model is used to compute single-hour sound values at individual sites. In contrast, multi-hour (site) values pertain to site-groups rather than individual hours. These site-group values are relevant when a model is used to compute site-group values averaged over many hours.

Table 28. Validation Matrix: Model Precision

Metric	Ambient sound levels used in computation	Computer model	Model random (rms) error		Correlation coefficient (ideal = 1.0)	
			Single hour (hourly)	Multi-hour (site)	Single hour (hourly)	Multi-hour (site)
Audibility	Measured	INM (A levels)	**12 %TmAud**	**12 %TmAud**	**0.7**	**0.6**
		INM (1/3 octaves)	**13 %TmAud**	**11 %TmAud**	**0.6**	**0.6**
		NMSIM	**9 %TmAud**	**6 %TmAud**	**0.8**	**0.9**
		NODSS	**10 %TmAud**	**3 %TmAud**	**0.7**	**0.94**
	EA	INM (A levels)	**17 %TmAud**	**15 %TmAud**	**0.3**	**0.2**
		INM (1/3 octaves)	**16 %TmAud**	**14 %TmAud**	**0.4**	**0.4**
		NMSIM	**12 %TmAud**	**8 %TmAud**	**0.7**	**0.8**
		NODSS	**9 %TmAud**	**5 %TmAud**	**0.8**	**0.92**
Equivalent Level (L_{eq})	————	INM (A levels)	**6 dB**	**4 dB**	**0.7**	**0.9**
		INM (1/3 octaves)	**6 dB**	**4 dB**	**0.7**	**0.9**
		NMSIM	**6 dB**	**3 dB**	**0.7**	**0.92**
		NODSS	**4 dB**	**5 dB**	**0.7**	**0.8**

8.6.3.1 Model random (rms) error

In Table 28, single-hour or hourly random errors for audibility range between 9 and 17%TmAud. In contrast, multi-hour or site values are lower, ranging between 3 and 15%TmAud. For equivalent level, single-hour random errors range between 4 and 6dB, while multi-hour values are lower,

ranging between 3 and 5dB. For any one row of the table, multi-hour values are less than single-hour values in all but two cases, because hour-to-hour discrepancies tend to average out.

8.6.3.2 Correlation coefficients

In Table 28, single-hour correlation coefficients range between 0.3 and 0.8, with all but two of them between 0.6 and 0.8. In contrast, multi-hour coefficients range between 0.2 and 0.92—a much wider range. Several of the models show extremely high multi-hour correlations—above 0.9.

For any one row of the table, multi-hour correlations are better than single-hour correlations in all but two cases. For example, audibility computations by NODSS, using measured ambients, have an hourly correlation coefficient of 0.7, but a much higher site-group correlation coefficient of 0.91. This improvement occurs because NODSS has a relatively modest site-group scatter about the average, as shown above in Figure 37 (lower-right panel), page 90. In contrast, audibility computations by INM (A levels), using measured ambient, have an hourly correlation coefficient of 0.7, but actually a lower site-group correlation coefficient of 0.6. This occurs because INM (A levels) has a large site-group scatter about the average, as shown in the upper-right panel of this same figure.

Both these measures of precision can be seen graphically in Figure 34 through Figure 39, pages 85 through 92, above. In these figures, more tightly clustered data points show lower scatter around their average, as well as higher correlation between measured and computed values.

8.7 Contour Error: Effect of Distance, Number of Averaged Hours, and Computed Metric

8.7.1 Overview

In Section 8.4, above, overall error was determined from "measurements versus computations," resulting in:

- Overall single-hour (hourly) error in Table 21, and
- Overall multi-hour (site) error in Table 22.

Overall errors of tour-aircraft sound contours are a mixture of these two sets of results, depending upon how many hours are averaged during contour computation. Where only one hour is computed, contour error matches overall single-hour error. In contrast, when a great number of hours are averaged into the contours, then contour error matches overall multi-hour error. In essence, averaging over many hours has reduced the random error in the contour computation. This section estimates contour error as a mixture of overall single-hour and overall multi-hour error.

It should be noted that this analysis of contour error was accomplished by examining the error at groups of sites, grouped by distance from the flight corridor. No contours were produced. In modeling, contours are normally produced by interpolation between and among specific points on the ground for which the computer model calculates the metric in question. Since the models in the study were used to compute this type of specific point data at many distances from the (Zuni Point) corridor, the errors (differences between computed and measured values) at these points, grouped by distance, can be used to estimate what error contours derived from these points would have.

Table 21 and Table 22, pages 89 and 93, provide only one error value for each model. That value is an average over all study data, both near and far from the flight track. But contour error depends

upon distance from the flight track. This section determines contour error as a function of this distance.

In brief, this section determines contour error by:

- Merging single-hour and multi-hour overall error, depending upon the number of hours averaged, and
- Analyzing the resulting merge as a function of distance from the flight track.

With this additional analysis, overall error can be determined anywhere on the sound contours computed by the study's models.

8.7.2 Averaging over many hours

The scatter in all "measured versus computed" plots is of two distinct types: partly site-to-site and partly hour-to-hour.

8.7.2.1 Site-to-site scatter

The *site-to-site* portion of the scatter is caused by site peculiarities that the models do not compute. Averaging over many hours cannot average out this part of the scatter. It is intrinsic to each site—that is, to each location on a computed sound-contour map. Therefore, it can intrinsically affect the contour error of each computer model.

A model that does not account for site peculiarities is less useful, since its sound contours have intrinsically more error. This site-to-site portion of the scatter limits the ultimate effectiveness of averaging over many hours. As more and more hours are averaged, contour error reduces towards a lower limit that is determined by the site-to-site scattter.

8.7.2.2 Hour-to-hour scatter

The *hour-to-hour* portion of the scatter is caused largely by hourly changes in wind and temperature gradients, or by other factors that change hourly but that the models do not compute. Hour-to-hour scatter can be reduced by averaging computation results over many hours, instead of just the one-hour intervals used in this study. Averaging over many hours would normally be done during use of the computer models for the Canyon. Since this part of the scatter can be "averaged out," it doesn't seriously limit the precision of the computer model.

- *Averaging equivalent level over many hours*. Hourly equivalent sound levels are energy-averaged to produce long-term equivalent levels. To compute corresponding multi-hour contours, it is most likely sufficient to average all input, in the normal manner, and then compute just once with this averaged input.

  Input averaging certainly accounts properly for hour-to-hour changes in air traffic. Acoustic energy is proportional to air traffic, and therefore air-traffic averaging will produce proper long-term equivalent sound levels.

  In contrast, however, averaging hour-to-hour meteorological input does not automatically guarantee proper long-term equivalent levels, especially at long propagation distances. Nevertheless, based upon insight from Section 9, below, meteorological input averaging appears to be reasonable for the Canyon.

In summary, input averaging for computation of equivalent level is most likely sufficient. Therefore, equivalent level contours should be computed in the following way:

- Average all the input values and then compute just once with this averaged input.

Such input averaging is common practice for equivalent-level contours.

- *Averaging audibility over many hours*. The situation is not nearly so straightforward for audibility averaging. Mathematically, multi-hour audibilities (%TmAud) are simple arithmetic averages of all the constituent hourly audibilities (%TmAud).

  However, input averaging will not work for this sound metric. Averaging air-traffic input for a single computation will likely give different results than computing each separate hour and then averaging the results. This is true because the relationship between air traffic and audibility is very "non-linear." When air traffic is doubled, audibility is not. For example, once audibility equals 100 percent, doubling air traffic cannot increase audibility further.

  Further analysis of the Canyon data could indicate how much error is introduced by input averaging. Until this analysis is performed, however, audibility contours should be computed the following way:

  - Compute separately for each hour, with its air-traffic and other input, and then
  - Average the results.

  A computer program that produces grid values for contouring, for example, would have to be run many times, the grid values averaged point by point, and then contoured.

Appendix I.2 presents graphical evidence of site effects that cannot average out.

8.7.3 Mathematical computation

To determine overall error as a function of distance, Eq.(1), page 84, was reused, but separately for data in the distance ranges from the corridor given in Table 29. These ranges were chosen so as to divide the total distance range into approximately equal bands.

Table 29. Distance Intervals Used for Contour Error Analysis and Associated Sites

Distance range from Corridor	Sites
1-to-3 miles	6All, 7All, 9CF
3-to-6 miles	4All, 5All, 8All
6-to-8 miles	3All
8-to-10 miles	2All
10-to-15 miles	1All, 9ABDE

Then these five overall errors, one per distance band, were plotted against distance.

As discussed just above, increasing the number of averaged hours will decrease the contour error, depending upon the split between hour-to-hour and site-to-site error. To determine this decrease, the variability in model discrepancy was analyzed to determine what part of it is hour-to-hour and what part is site-to-site. With this knowledge, graphs of contour precision were developed that show this dependence. Appendix I.3 provides further details.

8.7.4 Contour error: Audibility

8.7.4.1 Contour error graphs

Figure 56 and Figure 57 graph each model's audibility-contour error, depending upon whether these contours are computed with measured ambients or with EA ambients. Both these figures graphically show the dependence of contour error upon distance from the track, upon the number of hours that are averaged, and upon the computed value of audibility.

In the top frame of each figure, distance from the track PCA (point of closest approach) is plotted horizontally, while "variance" of audibility is plotted vertically.[57] The graph has four curves, one for each computer model. Inset in this top frame is a graph that provides a multiplier for variance, depending upon the number of hours that are averaged for the contours.

The lower frame of each figure converts the resulting variance into 95-percent confidence limits on the contour value, depending upon the computed value along the horizontal axis. In general, this conversion results in non-symmetrical error limits, because they must always lie between 0 and 100 percent.

8.7.4.2 Example

Figure 58 is an example use of Figure 56. The example circumstances appear at the top. The four example steps appear at the left edge and are illustrated with arrows and bold numbers to the right. The example's results appear at the bottom of the figure.

- *Step 1: Determine variance from distance*. In the upper-frame's main graph, draw a vertical line upward from the site's distance from track PCA (5 miles). Where this line hits the model's curve (NMSIM), turn it to the left to find the variance for a 100-hour average computation (25).
- *Step 2: Determine "multiplier" from number hours averaged*. In the upper-frame's embedded graph, draw a vertical line upward from the number of hours actually averaged (20), Where this line hits the model's curve (NMSIM), turn it to the left to find the multiplier (1.1). Note that all multipliers are equal to 1.0 for large number of averages—more than twenty-five averaged hours—the most common situation.
- *Step 3: Multiply variance by multiplier*. Multiply the variance (25) by the multiplier (1.1), to obtain 27. The bottom frame contains two sets of curves for these multiplied variances—one above the center, one below it. Interpolate between the labeled curves to approximate an upper and lower curve for 27.
- *Step 4: Determine upper and lower limit from computed %TmAud and the two curves from the multiplication of Step 3*. In the bottom frame, draw a vertical line upward from the computed %TmAud (25). Where this line hits the two interpolated curves (27), turn it to the left to find the upper and lower 95-percent confidence limits on the computed %TmAud (+10 and –9).

For this example, after averaging 20 hours of NMSIM computations using measured ambient, 25%TmAud contours at 5 miles have an error range between 16%TmAud and 35%TmAud, with 95-percent confidence. Figure 57 is used in the same manner.

[57] Variance is a specialized statistical term. It is needed here only to link the upper and lower graphs, as illustrated in the following section.

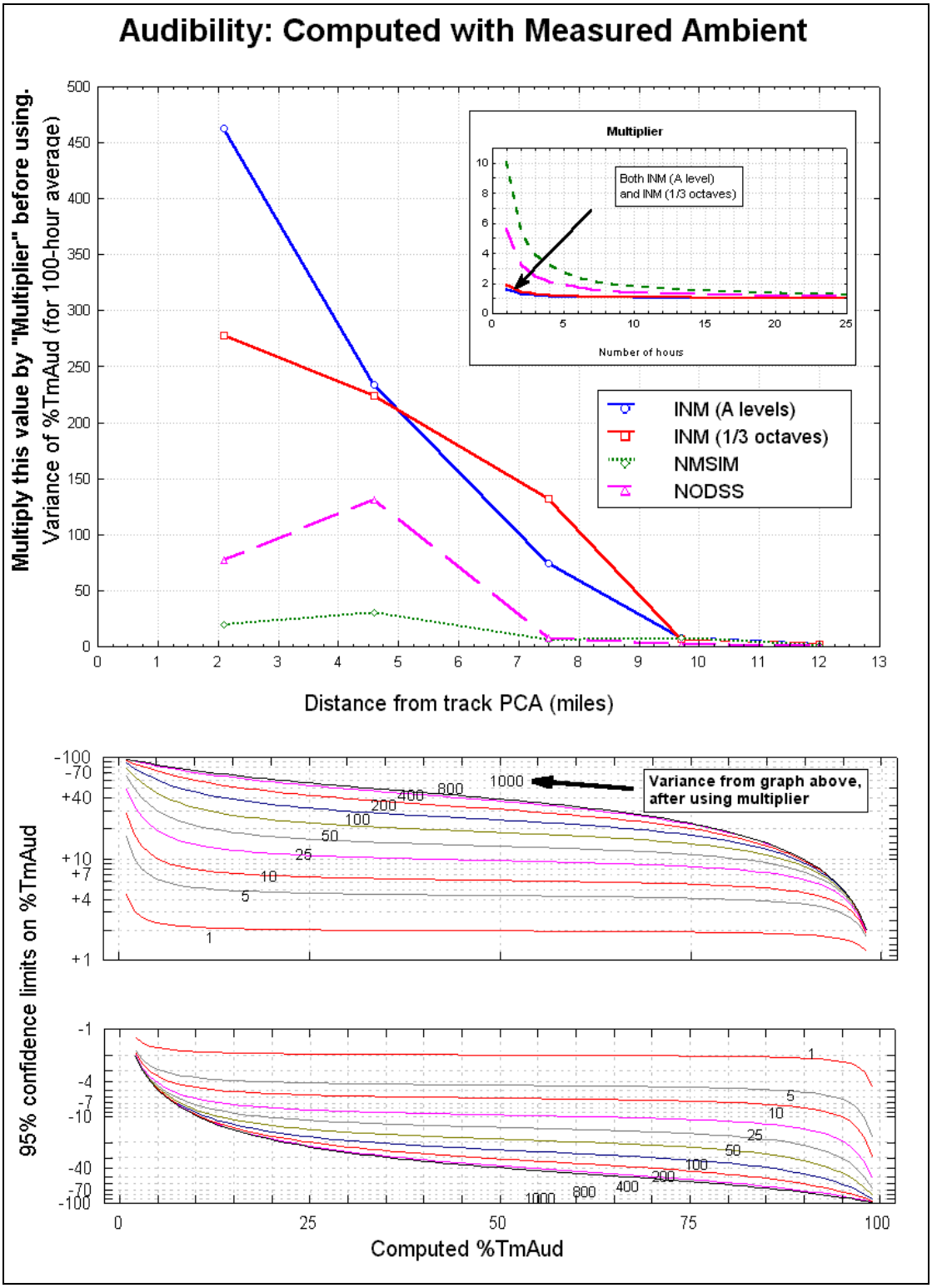


Figure 56. Confidence Limits for Audibility Contours (when Computed with Measured Ambients)

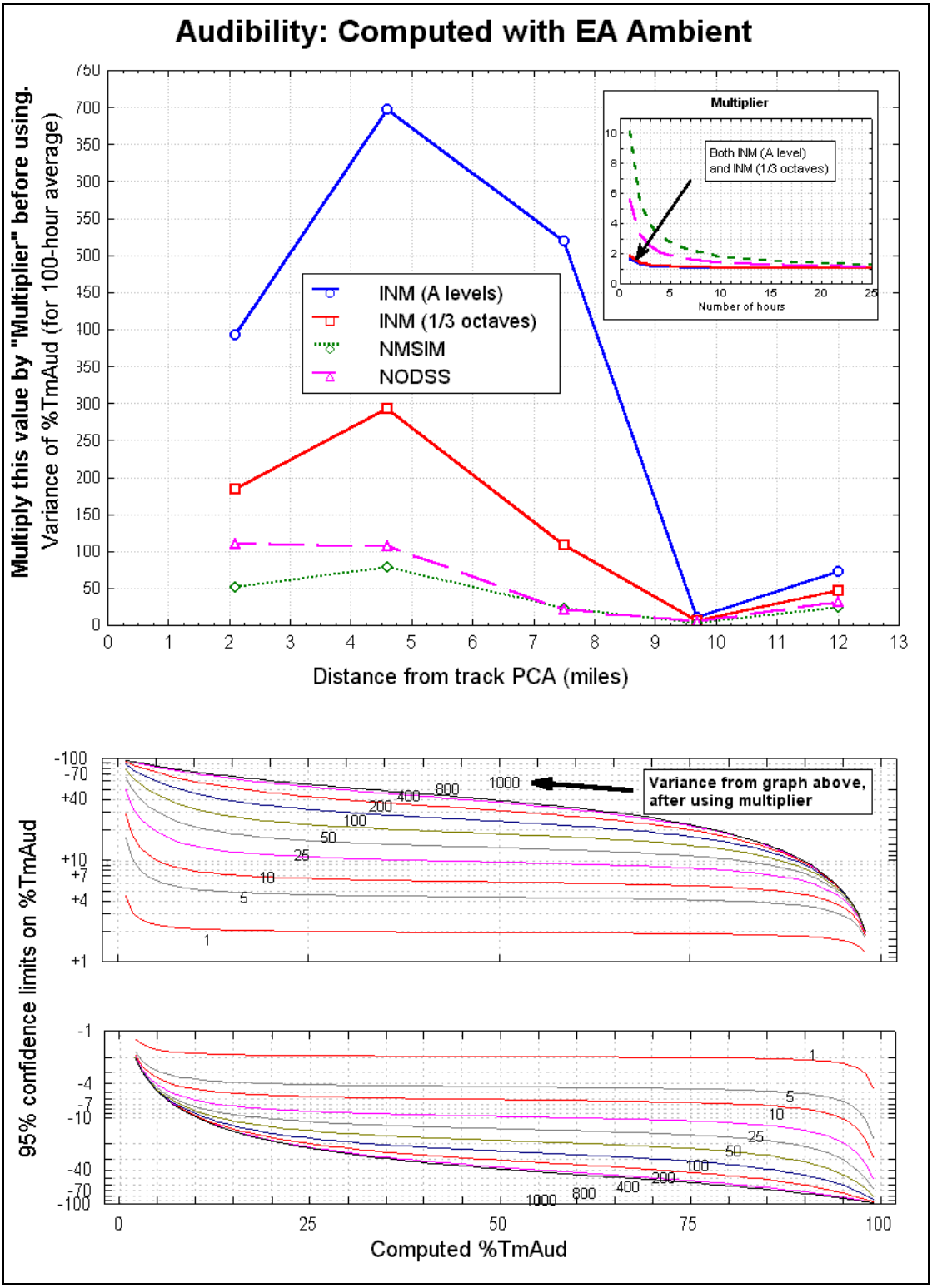


Figure 57. Confidence Limits for Audibility Contours (when Computed with EA Ambients)

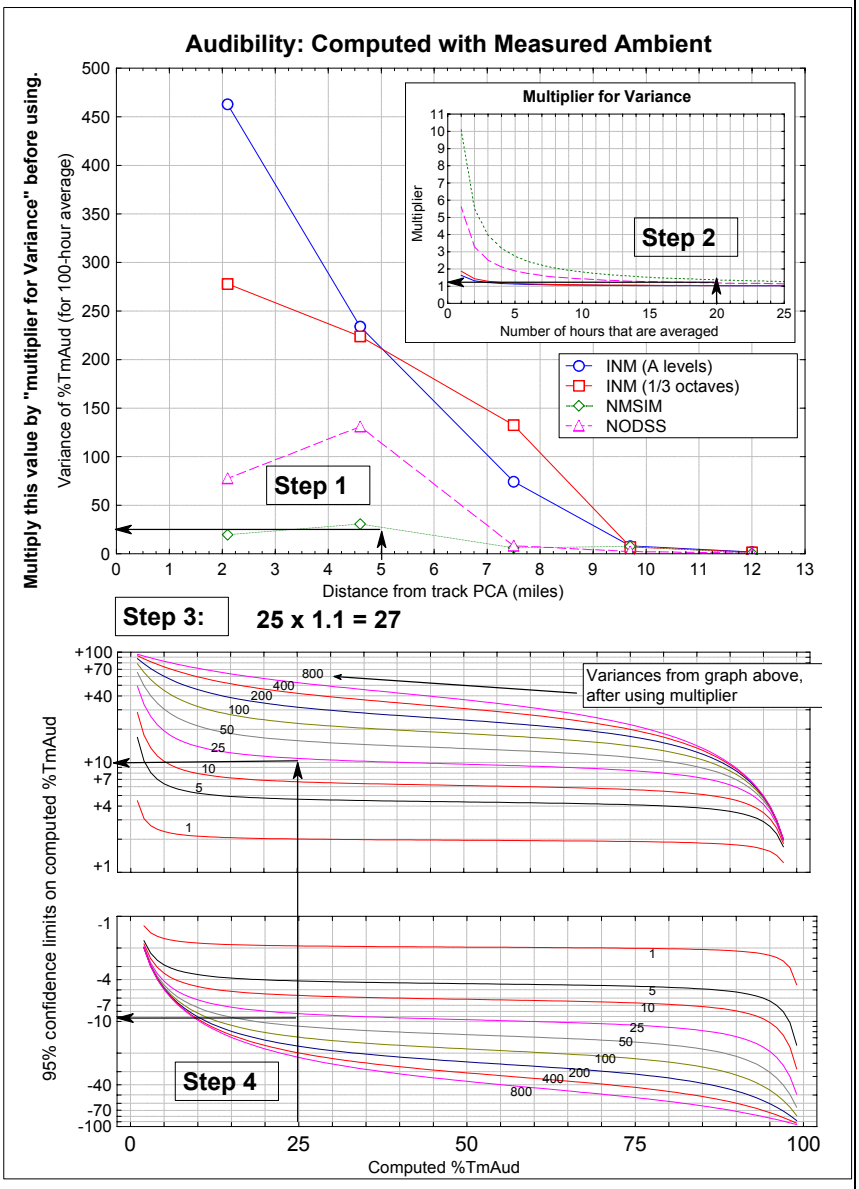
EXAMPLE
Audibility computed by NMSIM, using measured ambient and 20 averages.
Site is 5 miles from corridor
Site is on the 25%TmAud contour.

Step 1:
Determine variance from distance (5 miles yields variance of 25).

Step 2:
Determine "multiplier" from number hours averaged (20 avgs. yield mult. of 1.1).

Step 3:
Multiply variance by multiplier.

Step 4:
Determine upper and lower limit from computed %TmAud and the two curves from the multiplication of Step 3 (25%TmAud and value of 27 from Step 3 yield +10, -9).



RESULTING 95-PERCENT CONFIDENCE LIMITS
Upper limit = 25 + 10 = 35%TmAud.
Lower limit = 25 - 9 = 16%TmAud.

Figure 58. Example: Confidence Limits for Audibility Contours

8.7.4.3 General trends

The upper frames of Figure 56 and Figure 57 show the following general trends in audibility-contour error:

- Models with higher curves in the upper frame have larger contour errors, though the graph exaggerates the differences between models.[58]
- Though not shown in the graph, contour error below one mile or so is surely quite small. Then as flight-track distance increases, model error becomes greater—as expected. This pattern is obvious only for NMSIM and NODSS in the figure. In brief, the models have greater difficulty computing at larger distances because they do not account as well for larger-distance propagation factors.
- At some distance out from the flight track, contour error then starts to decrease. In spite of propagation factors, the models become better and better as audibility *reduces toward zero*. At extreme distances—50 miles, for example—all models would compute zero audibility, without any error. As a result of these general trends, these curves generally show an upward bulge, peaking where contour error is a maximum. These peaks occur between (approximately) 1 and 5 miles out from the flight track, depending upon model and ambient.

Most likely, the width and horizontal location of this vertical bulge depends upon air traffic. For air traffic less than in this study, large-distance zero audibility would occur closer in, causing the bulge to squeeze left in the figure, with reduced width. And vice versa: for more air traffic than in this study, the bulge would stretch to the right, with increased width. In either case, it would retain its general shape, as described in the three bullets above.

8.7.5 Contour error: Equivalent level

8.7.5.1 Contour error graphs

Figure 59 graphs each model's equivalent-level contour error. This figure graphically shows the dependence of contour error upon distance from the track and upon the number of hours that are averaged.

In the figure, distance from the track PCA (point of closest approach) is plotted horizontally, while 95-percent confidence limits are plotted vertically. The graph has four lines, one for each computer model. Inset in this graph is a smaller graph that provides a multiplier for these 95-percent confidence limits, depending upon the number of hours that are averaged for the contours.

[58] In particular, the upper frame's vertical scale is model "variance," which is roughly the square of model error. Therefore, taking square roots of the values on this scale gives a clearer numerical comparison between models. This square root is incorporated into the variance curves on the bottom graph, which are compressed downward.

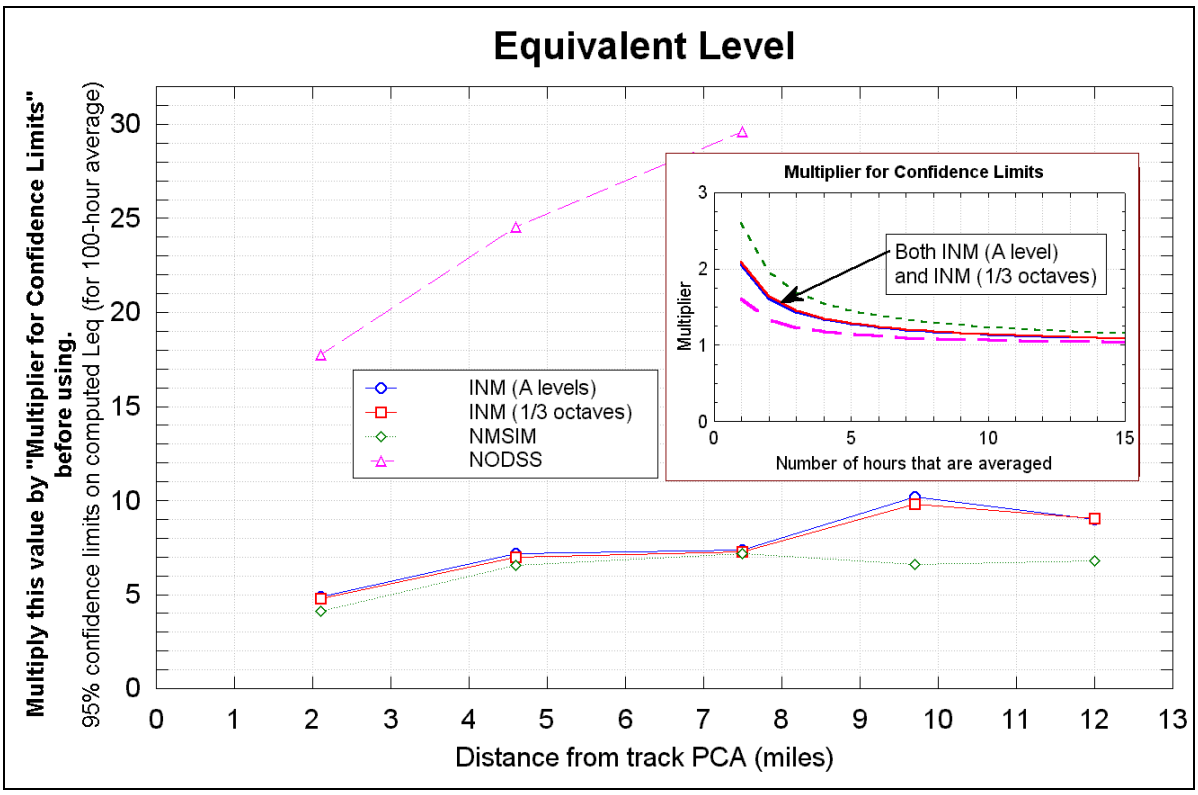


Figure 59. Confidence Limits For Equivalent Level Contours

8.7.5.2 Example

Figure 60 is an example use of Figure 59. The example circumstances appear at the top. The three example steps appear at the left edge and are illustrated with arrows and bold numbers to the right. The example's results appear at the bottom of the figure.

- *Step 1: Determine confidence limits (for 100-hour average) from distance*. Draw a vertical line upward from the site's distance from track PCA (5 miles). Where this line hits the model's curve (NMSIM), turn it to the left to find the confidence limits for a 100-hour average computation (6.5).
- *Step 2: Determine "multiplier" from number hours averaged*. In the imbedded graph, draw a vertical line upward from the number of hours actually averaged (30 was chosen because for this large number, the multiplier should be approximately 1, even though 30 is beyond the limits of the graph). Where this line hits the model's curve (NMSIM), turn it to the left to find the multiplier (1.0). Note that all multipliers are equal to 1.0 for large number of averages—the most common situation.
- *Step 3: Multiply 100-hour confidence limits (from Step 1) by multiplier (from Step 2)*. Multiply the confidence limits (6.5) by the multiplier (1.0), to obtain 6.5.

For this example, after averaging 30 hours of NMSIM computations, contours at 5 miles have an error range of ±6.5 dB. In particular, a 20dB contour ranges from 13.5 to 26.5 dB, with 95-percent confidence.

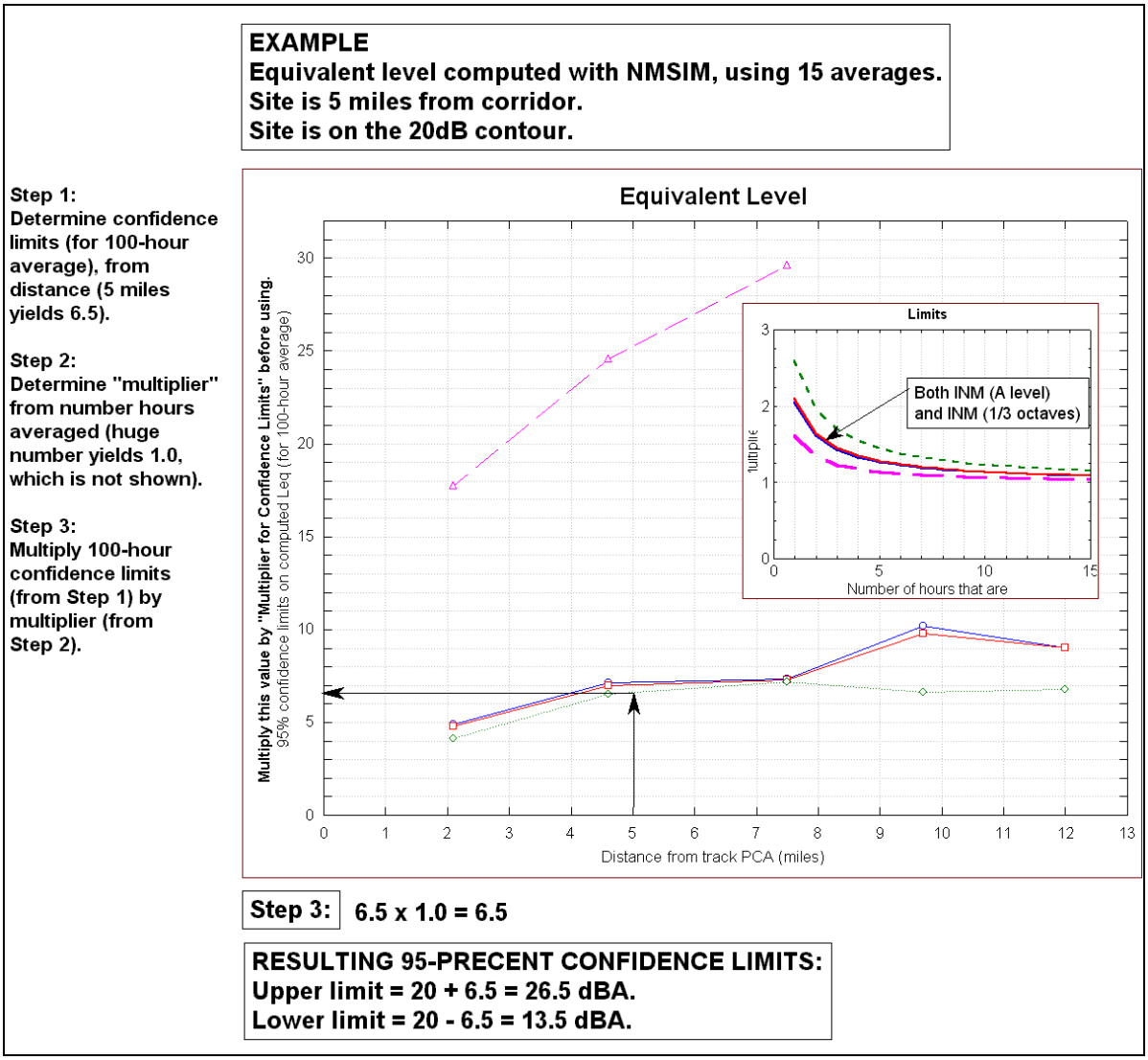


Figure 60. Example: Confidence Limits for Equivalent Level Contours

8.7.5.3 General trends

Figure 59 shows the following general trends in equivalent-level contour error:

- Models with higher curves have larger contour errors. In particular, NODSS shows a very large contour error, corresponding to its very large overall errors in Table 21 and Table 22, above.
- As flight-track distance increases from zero, model error becomes greater, as expected. In brief, the models have greater difficulty computing at larger distances because they do not account as well for larger-distance propagation factors. This "distance penalty" levels off around 5-to-6 miles from the flight track for NMSIM, but not for either version of INM.

9. INSIGHTS ABOUT MODEL DISCREPANCIES

9.1 Overview

This section contains a partial diagnosis of each model's overall error, accuracy and precision. This partial diagnosis may provide some insight into potential reasons for the model discrepancies of Section 8.

- Section 9.2 shows the association between *model discrepancies* and physical factors that were simultaneously measured in the field.
- Section 9.3 shows the association between *measured audibility* and these same physical factors. Most importantly, this part of the analysis directly measures the effectiveness of Vistaliners in reducing aircraft audibility.
- Section 9.4 contains a *direct comparison among the model computations*, each model with the other.

9.2 Model Discrepancies versus Physical Factors

This section shows the association between model discrepancies and physical factors that were simultaneously measured in the field—that is, model discrepancies versus physical factors. This partial diagnosis of model discrepancies may prove useful if the decision is made to improve model performance.

This analysis looks for "associations" between discrepancies and various physical factors—some incorporated in the computer models, some not. It does not say that any particular factor "causes" model discrepancy, only that changes in that factor occurred in tandem with changes in discrepancy. In words, when the physical factor increased in value, the discrepancy consistently increased, as well. Or vice versa—when the physical factor increased, the discrepancy consistently decreased.

The associations that result from this analysis therefore have to be examined together with possible physical explanations, to gain insight into their meaning—as is done in Section 9.2.4. In brief, interpretation of these results needs acoustical reasoning in addition to these numerical associations.

Note that in analyzing audibility discrepancies, only computations using measured ambient sound levels were examined.

9.2.1 Method

A great number of physical factors potentially underlie model discrepancies. Simple plots of "model discrepancy versus one physical factor at a time" may not always be useful for diagnosis, for the following reason. Each plot contains only one physical factor, out of a large number that were simultaneously measured. Although this plot may show some relationship between its physical factor and model discrepancy, that relationship can sometimes be influenced by the physical factors not in the plot. In extreme situations, the missing factors can possibly distort the pattern of points on the plot—thereby giving it an apparent shape when it truly has none, or thereby masking the true relationship between model discrepancy and that plot's factor.

In brief, graphical diagnosis is limited to one factor at a time and can possibly be misleading for that reason. The interrelationships between discrepancy and physical factors are sometimes just too complicated for simple graphical diagnosis.

In contrast, multiple linear regression is sometimes capable of simultaneously sorting out the complicated relationship between model discrepancy and all the physical factors, in one mathematical operation. Such regression of "model discrepancy versus all the physical factors at the same time" results in a regression coefficient for each factor. These regression coefficients directly indicate the apparent association of that factor with model discrepancy, taking into account the simultaneous associations of all the other factors, as well.

The mathematical techniques for multiple linear regression are well standardized. For this study, the computer program Statistica[59] was used. The same results would follow had a different program been used. All physical factors in the data were initially included in each regression. Then those factors were omitted that had less than 90-percent chance of being important to the regression. Interaction terms were not tested. "Distance to the track" and "visible angle" were moderately correlated (correlation coefficient = –0.65). Rather than just retaining one of them, they were both retained in the regression whenever either one seemed important to the regression.

9.2.2 Numerical results

Each multiple linear regression results in an equation of this type:

$$\begin{aligned} \text{Discrepancy} &= S_{computed} - S_{measured} \\ &= \text{Constant} + C_1F_1 + C_2F_2 + C_3F_3 + \text{L} \quad , \end{aligned} \tag{6}$$

where F represents each physical factor, and C represents that factor's regression coefficient. The constant is the residual discrepancy not associated with any of the factors. By substituting specific values of these physical factors into this equation, we can compute the association between those factors and model discrepancy, $S_{computed} - S_{measured}$. Table 30 through Table 32 show the results of these multiple linear regressions and the following text describes the table and its information in detail.

- The first column in each table contains all the physical factors that proved important to one or another of these regressions.
- The second column shows the measured range of that physical factor, from that factor's minimum values to its maximum value. The range for "angle of visibility" (Table 32) is shown reversed, from "maximum" to "minimum." It is reversed because the maximum visible angle (165 degrees) occurs at sites with the minimum distance (8,943 feet). This reversal simplifies the interpretation of the regression results in Section 9.2.3.
- The third column identifies the computer model.
- The fourth column shows the resulting regression coefficient.

[59] *Statistica for Windows (Computer Program Manual),* StatSoft Inc., www.statsoft.com, Tulsa OK, 1999

Table 30. Associations between Model Discrepancies and Physical Factors (1 of 3)

Physical factor	Factor range	Model	Regression coefficient	Corresponding association with model discrepancy (notes 1 and 2)	Summary of numerical associations (audibility computed with measured ambients)
Temperature gradient	–5.1 to –1.3 Fdeg/1000ft	INM A	+2.30	–12 to –3 %TmAud	All models moderately overcompute equivalent level (positive values), with the largest overcomputations (+5dB, +5dB, +4dB, +9dB) during the most extreme temperature gradients (–5.1 Fdeg/1000ft). Just the opposite is true for audibility—that is, the models all undercompute. These associations are quite consistent from model to model.
		INM 1/3	+0.897	–5 to –1 %TmAud	
		NMSIM	+2.50	–13 to –3 %TmAud	
		NODSS	+3.48	–18 to –5 %TmAud	
		INM A	–1.01	+5 to +1 dB	
		INM 1/3	–0.953	+5 to +1 dB	
		NMSIM	–0.880	+4 to +1 dB	
		NODSS	–1.68	+9 to +2 dB	
Component of wind from track PCA to site	–8 to +12 kt	INM A	+0.976	–8 to +12 %TmAud	When wind blows hardest from site to track (–8 kt) (upwind propagation), most models overcompute (+9%, +8%, +7dB, +7dB, +7dB, +3dB). And vice versa, when wind blows hardest from track to site (+12 kt) (down wind propagation), most models undercompute (–14%, –12%, –10dB, –10dB, –10dB, –4dB). However, both versions of INM do just the opposite for audibility. In addition, these values center around zero, for this factor that can change hourly.
		INM 1/3	+0.451	–4 to +5 %TmAud	
		NMSIM	–1.14	+9 to –14 %TmAud	
		NODSS	–0.971	+ 8 to –12 %TmAud	
		INM A	–0.872	+7 to –10 dB	
		INM 1/3	–0.863	+7 to –10 dB	
		NMSIM	–0.842	+7 to –10 dB	
		NODSS	–0.326	+3 to –4 dB	
Component of wind along flight track	–6 to +7 kt	INM A	+2.05	–12 to +14 %TmAud	The signs of all audibility discrepancies match the sign of the wind component along the track. For equivalent level, the match is generally of opposite sign. In addition, these values center on a value of zero, for this factor that can change hourly.
		INM 1/3	+4.70	–28 to +33 %TmAud	
		NMSIM	+1.13	–7 to +8 %TmAud	
		NODSS	+2.32	–14 to +16 %TmAud	
		INM A	–0.581	+3 to –4 dB	
		INM 1/3	–0.677	+4 to –5 dB	
		NMSIM	–0.753	+5 to –5 dB	
		NODSS	+0.758	–5 to +5 dB	
Wind speed	0 to 12 kt	INM A	+1.32	0 to +16 %TmAud	All tabulated models overcompute the harder the wind is blowing, independent of wind direction. This association is in addition to the wind-component associations.
		INM 1/3	+1.34	0 to +16 %TmAud	
		NMSIM	+0.587	0 to +7 %TmAud	
		NODSS			
		INM A	+0.495	0 to +6 dB	
		INM 1/3	+0.455	0 to +5 dB	
		NMSIM			
		NODSS	+0.444	0 to +5 dB	
Atmospheric absorption at 200 Hz	0.236 to 0.423 dB/1000ft	INM A	+232	–19 to +25 %TmAud	These values center around a value of zero, for this factor that can change hourly
		INM 1/3	+133	–11 to +14 %TmAud	
		NMSIM			
		NODSS			
		INM A	–44.0	+4 to –5 dB	
		INM 1/3	–46.0	+4 to –5 dB	
		NMSIM	–52.7	+4 to –6 dB	
		NODSS	–35.6	+3 to –4 dB	

Note 1. Positive discrepancy means the model *overcomputes*; negative means it *undercomputes*. Empty cells mean the regression did not find a reliable value; the actual value may be large, nevertheless.

Note 2. Discrepancy associations were forced to equal zero for the following: zero temperature gradient, zero wind, and atmospheric absorption produced by computer-input values of temperature and relative humidity.

Table 31. Associations between Model Discrepancies and Physical Factors (2 of 3)

Physical factor	Factor range	Model	Regression coefficient	Corresponding association with model discrepancy (notes 1 and 2)	Summary of numerical associations (audibility computed with measured ambient)
Broadband L50 ambient	15 to 36 dB	INM A	–1.68	+15 to –20 %TmAud	These values center around zero, for this factor that can change hourly (relative to its average measured value for each half-day period).
		INM 1/3	–0.387	+3 to –5 %TmAud	
		NMSIM			
		NODSS	+0.908	–8 to +11 %TmAud	
		INM A	–0.298	+3 to –4 dB	
		INM 1/3	–0.309	+3 to –4 dB	
		NMSIM	–0.253	+2 to –3 dB	
		NODSS	+0.292	–3 to +4 dB	
Ambient: water		INM A	–22.2	–22 %TmAud	Sites with water-related ambient had moderate undercomputation of audibility – for all models except NMSIM.
		INM 1/3	–15.9	–16 %TmAud	
		NMSIM			
		NODSS	–8.41	–8 %TmAud	
		INM A			
		INM 1/3			
		NMSIM			
		NODSS	–3.31	–3 dB	
Ambient: coniferous forest		INM A	–6.78	–7 %TmAud	Coniferous forest sites had moderate overcomputation (except for INMA) - larger for audibility than for equivalent level.
		INM 1/3	+15.9	+16 %TmAud	
		NMSIM	+26.8	+27 %TmAud	
		NODSS	+18.6	+19 %TmAud	
		INM A	+4.36	+4 dB	
		INM 1/3	+3.98	+4 dB	
		NMSIM	+5.38	+5 dB	
		NODSS			
Ambient: pinyon juniper		INM A	+5.37	+5 %TmAud	Pinyon juniper sites also had moderate overcomputation of audibility, but much less than for coniferous forest sites.
		INM 1/3	+8.48	+8 %TmAud	
		NMSIM	+8.09	+8 %TmAud	
		NODSS			
		INM A			
		INM 1/3			
		NMSIM			
		NODSS			

Note 1. Positive discrepancy means the model *overcomputes*; negative means it *undercomputes*. Empty cells mean the regression did not find a reliable value; the actual value may be large, nevertheless.

Note 2. Values were forced to equal zero for the following: average measured broadband L50 ambient, and desert scrub ambient vegetation zone so other zone's values are relative to "desert scrub."

Table 32. Associations between Model Discrepancies and Physical Factors (3 of 3)

Physical factor	Factor range	Model	Regression coefficient	Corresponding association with model discrepancy (notes 1 and 2)	Summary of numerical associations (audibility computed with measured ambient)
Perpendicular distance to the track PCA	8943 to 78248 ft	INM A	–0.0007	0 to –49 %TmAud	In general, models undercompute at increased distances and overcompute at decreased angles of visibility. When combined (lowest set of numbers to the left), these two trends partially cancel one another out. For audibility, combined overcomputation for INM (1/3 octaves) and NMSIM is large. For equivalent level, undercomputation for NODSS is large.
		INM 1/3	–0.0002	0 to –14 %TmAud	
		NMSIM	0.0000	0 to 0 %TmAud	
		NODSS			
		INM A	–0.00014	0 to –10 dB	
		INM 1/3	–0.00017	0 to –12 dB	
		NMSIM	–0.000064	0 to –4 dB	
		NODSS	–0.00056	0 to –39 dB	
Angle of visibility	165 to 14 deg	INM A	–0.291	0 to +44 %TmAud	
		INM 1/3	–0.301	0 to +45 %TmAud	
		NMSIM	–0.18	0 to +27 %TmAud	
		NODSS			
		INM A	–0.123	0 to +19 dB	
		INM 1/3	–0.124	0 to +19 dB	
		NMSIM	–0.0767	0 to +12 dB	
		NODSS	–0.0045	0 to +1 dB	
Distance and angle, combined	(same as above)	**INM A**	———	**0 to –5 %TmAud**	
		INM 1/3	———	**0 to +31 %TmAud**	
		NMSIM	———	**0 to +27 %TmAud**	
		NODSS			
		INM A	———	**0 to +9 dB**	
		INM 1/3	———	**0 to +7 dB**	
		NMSIM	———	**0 to +8 dB**	
		NODSS	———	**0 to –38 dB**	

Note 1. Positive discrepancy means the model *overcomputes*; negative means it *undercomputes*. Empty cells mean the regression did not find a reliable value; the actual value may be large, nevertheless.

Note 2. Discrepancy associations were forced to equal zero for the following parameter values: minimum perpendicular distance (where propagation is best computed), and maximum angle of visibility (same reason).

- The fifth column shows the factor's numerical associations with model discrepancy. Each association appears as a range of values—for example, –12 to –3%TmAud (first entry, Table 30). This range of values matches the factor's range. For example, the value of –12 occurs for temperature gradients of minus 5.1Fdeg/1000ft (from the second column), while the value of –3 occurs for temperature gradients of –1.3Fdeg/1000ft. These values are computed from that factor's regression coefficient.
  - Where cells are empty in this column, the regression did not find a reliable value. That factor's regression coefficient was not important to the regression, generally because the association is small or zero.[60]
  - Where these values center around zero for factors that can change hourly, they would tend to "average out" from hour to hour. For this reason, averages over many hours would not be influenced by these factors. This is especially likely if the numerical associations are small, as well.

[60] It is possible, however, that the effect might be large but that large data scatter makes it impossible to discover with this mathematical technique. In addition, perhaps a factor's effect is not related *linearly* to the factor. In that case, linear regression might not detect the effect, either.

- The last column summarizes the numerical associations, for interpretation in the following section.
- In the tables, Note 1 should be self-explanatory. Note 2 addresses a technique that was used to help make the associations more meaningful. The range of the associations (from minus to plus or *vice versa*) is determined by the regression. Where these ranges fall, can be selected by how the analysis is done. For the parameters listed in Note 2 on each table, the zero value of the range was forced by normalization to be associated with a specific value of the parameters listed. For example, the ranges were forced to equal zero when the temperature gradient was zero since the models do not account for temperature gradient and were assumed therefore to have the least error when the gradient was zero.

9.2.3 Interpretation of these numerical results

This section interprets the numerical results of Table 30 through Table 32, relying upon the summaries in the right-most column of these three tables.

9.2.3.1 Hourly factor: Temperature gradient

The regression shows moderate *over*computation of equivalent level, which is expected from sound-propagation theory. In brief, negative temperature gradients produce upward refraction, which attenuates measured sound levels. This temperature-gradient attenuation is not computed by any of the models, so they would be expected to overcompute. This moderate overcomputation is consistent with conclusions in Section 9.3, below—where *measured* sound-level attenuation at larger distances is associated with negative temperature gradients.

In contrast, audibility *under*computation is not expected from sound propagation theory. It is possible that the regression is blending this temperature-gradient effect with the effect of increasing distance from the flight track (certainly temperature-gradient attenuation increases with increasing distance). See Section 9.2.3.7 for further discussion of this possibility.

9.2.3.2 Hourly factors: Both wind components

*Component from track to site. Over*computation is expected from basic acoustics, when wind blows from site to track (upwind propagation). In brief, upwind sound propagation produces upward refraction, which attenuates measured sound levels. This wind-induced attenuation is not computed by any of the models, so they would be expected to overcompute. In a similar manner, *under*computation is expected from basic acoustics, when wind blows in the opposite direction (downwind propagation). Downward refraction increases measured sound levels (especially when terrain intervenes), and the models ignore this effect.

For equivalent levels, all models exhibit this expected association. For audibility, only NMSIM and NODSS exhibit this expected association. In contrast, both these apparent effects are just the opposite for both INM versions, for unknown reasons.

Component along track (perpendicular to the component just discussed). The signs of all audibility discrepancies match the sign of the wind component along the track. This indicates that wind in the direction of aircraft travel is associated with relatively large *over*computation. The opposite is true when the wind shifts by 180 degrees (*under* computation when wind is opposite the direction of aircraft travel). For equivalent level, these values are less, and also of opposite sign.

Basic acoustics does not suggest which algebraic sign is expected, for either metric.

Both components. These wind values center numerically around zero, for this factor that can change hourly. Therefore, wind-related differences between computation and measurement will tend to average out from hour to hour. For this reason, average computations over many hours should not be influenced by wind. The result that the models tend to overcompute levels when wind blows toward the site, and undercompute when wind blows toward the corridor is consistent with conclusions in Section 9.3, below—where measured sound-level attenuation at larger distances is affected similarly by wind direction.

9.2.3.3 Hourly factor: Wind speed

All models *overcompute* both metrics the harder the wind is blowing, independent of wind direction. This wind-speed association is in addition to the wind-component associations just discussed.

Two possibilities exist for this overcomputation. The first possibility seems plausible, but is probably not correct. It states that this type of effect for audibility is expected from basic acoustics. In brief, as wind speeds pick up, ambient noise increases and so measured audibility is reduced. Because the models do not include this audibility reduction, they would tend to overcompute. This possible explanation would not apply to equivalent level, however, since it is independent of ambient sound level.

The reason this possible explanation is probably not correct concerns the wind speed that was used to normalize the linear regression—the regression math was normalized to zero wind. We did this because none of the models include wind speed, thereby causing these numerical associations to relate to model input. However, model input does include ambient sound levels, which are related to wind. The "average" wind over each half-day period produced the average measured ambient input for the model computations.

On the other hand, if we had normalized to average wind speed, instead of zero, then the tabulated values would center numerically around zero, for this factor that can change hourly (up and down from its average, half-day value). For this reason, average computations over many hours should not be influenced by wind speed—in the same manner as for wind components. This explanation is valid, we believe, and applies to both sound metrics.

9.2.3.4 Hourly factor: Atmospheric absorption

These values center numerically around zero, for this factor that can change hourly (up and down from the value used for the input temperature and input relative humidity). Therefore, changes in hour-to-hour sound propagation, due to hourly changes in atmospheric attenuation, will tend to average out. For this reason, average computations over many hours should not be influenced by hourly changes in atmospheric attenuation.

9.2.3.5 Site factor: Broadband[61] L_{50} ambient

These values also center numerically around zero, for this factor that can change hourly (relative to the input half-day-average input). Therefore, changes in hour-to-hour audibilities, due to hourly changes in broadband L_{50} ambient, will tend to average out. For this reason, average computations over many hours should not be influenced by hourly changes in ambient sound level.

[61] "Broadband" is used here to distinguish from "⅓ octave band" data and analysis. Broadband simply means "A-weighted".

Note that these numerical results were centered around zero for L_{50} because L_{50} was the value used to define the ambients used in modeling. Thus it should result, logically, in the lowest discrepancy. Because L_{50} is exceeded half the time, by definition, with its use the models should tend to undercompute audibility half the time and overcompute it the other half.

9.2.3.6 Site factor: Ambient vegetative zones, compared to desert scrub

Water-dominated ambient. Water-dominated ambient is associated with moderate *under*computation of audibility, except for NMSIM—for unknown reasons.

Coniferous forest and pinyon juniper ambient. Coniferous forest is associated with moderate *over*computation, perhaps because the models do not account for acoustic shielding by trees that break lines-of-sight over long distances.

This overcomputation is larger for audibility than for equivalent level, perhaps because audibility depends mostly on aircraft at their fringes of audibility (extreme distances), while equivalent levels depend mostly upon aircraft at their closest point of approach. Therefore, critical propagation distances are larger for audibility, causing more tree shielding.

Pinyon juniper is also associated with moderate overcomputation, but much less than coniferous forest. This is consistent with pinyon juniper's far sparser vegetation. Note that propagation distances through this sparse vegetation were extremely large for some study sites.

9.2.3.7 Site factors: perpendicular distance and angle of visibility

The regression results for perpendicular distance and for angle of visibility must be considered together, because these two factors are correlated with each other in this study. When distance is low—nearby sites—visibility angle is large, and vice versa. For this reason, the sum of these two factors' numerical values is more physically meaningful than is either factor alone.

In general, the models *undercompute* at increased distances and *overcompute* at decreased angles of visibility. When combined (bold entries in Table 32), these two numerical values partially cancel one another out. In particular:

- *NODSS equivalent level.* NODSS computation of equivalent level is an anomaly here. Most likely, this anomaly is connected with the very large NODSS undercomputation of equivalent level (see Figure 36 and Table 22, above). This current analysis seems to suggest that NODSS is miscomputing the effect of propagation distance. A detailed look into NODSS computations is necessary here.
- *INM (A level) audibility.* INM (A level) computation of audibility is also an anomaly here, compared to the other models. On the surface, INM (A level) seems like the only model that doesn't miscompute the distance-angle effect. However, something else appears to be happening, based upon the other models—something that INM (A levels) somehow misses—as described in the following bullet.

  In addition, neither INM model computes shielding due to terrain. It may be that this omission is showing up here in the analysis—though only for INM (A levels) and not for INM (1/3 octaves), which also omits terrain shielding.
- *All other situations, possibly caused by temperature gradient.* In all other situations, the models *overcompute* significantly at large distance (small angles). This overcomputation at large

distances may be an actual effect of temperature gradients, rather than distance. Other studies show that negative vertical temperature gradients cause very large sound-level reductions during daytime, at these large distances, which would result in very large model overcomputations (since the models ignore this effect). Perhaps the linear regression assigned this overcomputation to "distance." Because this overcomputation is physically caused by vertical temperature gradients, it becomes strongly associated with increased distance in this analysis.

9.2.4 Model insights from this analysis

The following model insights follow from this analysis:

- *Broadband L50 ambient.* Choice of L_{50} as the metric for determining broadband ambient is probably important here.`
- *Vertical temperature gradients.* The combined effects of distance, angle of corridor visible and vertical temperature gradient cannot be easily separated. However, for equivalent levels, we suspect that most of the association between discrepancies and distance from the flight track, as well as visible angle, is actually caused by vertical temperature gradients. Lack of such algorithms for this acoustical factor may be causing overcomputation, at least of equivalent levels, in all models at the largest distances in this study (10-to-15 miles).

  Adding the average temperature gradient's values (Table 30) to the combined values for distance and angle (Table 32, bold entries) yields association ranges of:

  - INM A: –7 to –12 %TmAud,
  - INM 1/3: –3 to +28 %TmAud,
  - NMSIM: –8 to +19 %TmAud,
  - NODSS: –18 to –5 %TmAud,
  - INM A: +3 to +12 dB,
  - INM 1/3: +3 to +10 dB,
  - NMSIM: +2 to +10 dB, and
  - NODSS: +5 to –33 dB (general difficulty with NODSS's equivalent level as noted).

  The predominant pattern in these numbers, at least for equivalent levels, is an overcomputation, especially if NODSS results are ignored.
- *Other meteorological algorithms.* Inclusion of other meteorological algorithms seems not necessary, for computations averaged over many hours. In particular:
  - Inclusion of average temperature and relative humidity, to include average atmospheric absorption, appears adequate.
  - Omission of wind algorithms appears adequate.
- *Tree shielding.* To accurately compute either sound metric, all models may require algorithms to compute shielding from large expanses of intervening trees. Lack of such algorithms may be causing overcomputation by approximately 16-to-27%TmAud and 4-to-5 dB, on the average over the distances in this study (2-to-15 miles). This need is greatest for coniferous forest, less so for regions of pinyon juniper.
- *NODSS equivalent level.* NODSS is greatly undercomputing equivalent level—as is apparent in preceding sections of this report.
- *INM (A level) audibility.* This analysis suggests that INM (A level) is miscomputing in some manner associated with distance or angle of visibility. This suggestion follows from INM's

computations relative to all the other models, rather than from direct regression evidence against INM, itself. See Section 9.3 for further details.

A general caveat: Multiple linear regression is an inexact method of identifying possible reasons for model discrepancies, since it forces a linear form on all discrepancy-factor relationships. Such a functional form can sometimes produce phantom relationships and can sometimes hide real relationships. Nevertheless, this analysis should provide some useful guidance towards future model improvements.

9.3 Measured Audibility versus Physical Factors

This section shows how measured audibility depends upon physical factors that were simultaneously measured in the field—that is, measured audibility versus physical factors. This analysis may also provide insight into future model improvements, to supplement the insights of Section 9.2.

For example, this analysis shows that Vistaliners are significantly less audible than other aircraft types, all else being equal. Such knowledge may influence NPS and / or FAA policy concerning "quiet technology" aircraft. As another example, this analysis shows that wind and temperature gradients are not important to the measured tour-aircraft audibility. Therefore, lack of these algorithms in the model computations probably doesn't contribute to model discrepancies.

This analysis also yields an empirical relationship—a very long equation—that computes audibility from the same input used by the computer models. This empirical relationship was first intended as a Failsafe Method (see Section 2.2.3 above), to estimate tour-aircraft audibility in case none of the computer models proved sufficiently accurate. Its use in that capacity is not considered necessary, since the models tested are deemed suitable for use under specified circumstances (see Sections 10 and 11).

9.3.1 Method

Non-linear regression was used to gain possible insight into the physical factors that may affect measured tour-aircraft audibility, including the magnitude of their effect. In brief, this non-linear regression combines measured audibility and all these factors into a single mathematical equation that computes audibility from these physical factors. This equation mimics several of the algorithms within the computer models, but much simplified. It is based upon known acoustical principles, but leaves some leeway for adjustment (regression coefficients) to best match the data.

APPENDIX J, page 243 contains further details about this non-linear regression method. APPENDIX K, page 245 contains the full input to this regression, while APPENDIX L, page 255 contains the resulting regression equation.

9.3.2 Numerical results

Figure 61 shows the end result of this non-linear regression. Plotted horizontally in the figure is the audibility for each site-hour, computed with the best-fit regression equation. Plotted vertically are the corresponding measured audibilities.

As the figure shows, the regression fit is quite tight—actually somewhat better than the fit from some of the computer models in Figure 41 and Figure 42, above. However, regression models of this type always fit measurements well, because they are actually derived from those measurements. If this

regression equation were used to predict future values and then compared to future measurements, the fit would not be this tight.

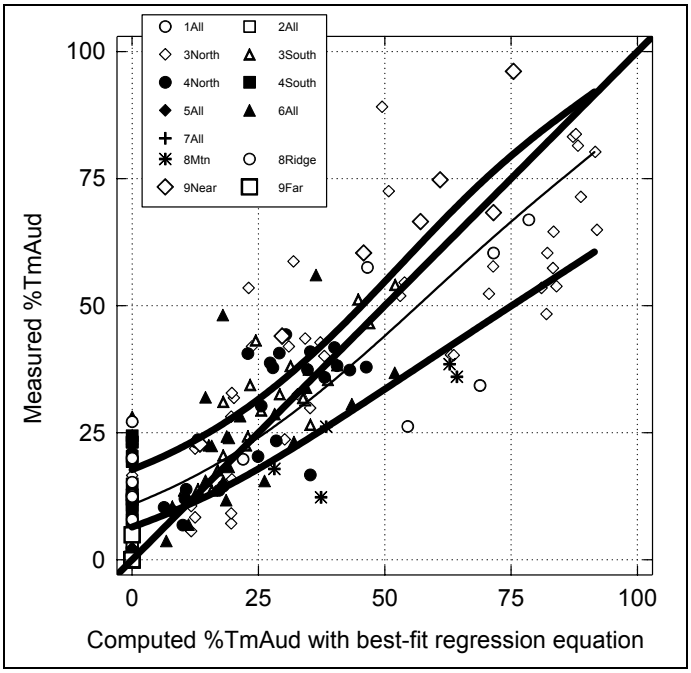


Figure 61. Best Regression Fit to Measured %TmAud

Table 33 shows the computed results from this non-linear regression:

- The first column contains the physical factor of interest.
- The second column summarizes the association with this physical factor, according to this empirical relationship.

Table 33. Association of Physical Factors with Measured Audibility

Physical factor	Association of this physical factor with measured audibility (from magnitude of the regression coefficients)
Percentage of Vistaliners	Vistaliners produced only 0.3 times the audibility that non-Vistaliners produce, for equal numbers of each per hour.
Terrain shielding	Terrain shielding averaged 13 dB for all measurements.
Wind speed and direction, combined with terrain shielding	When aircraft were shielded by terrain, 10-knot upwind propagation (site to track) increased terrain shielding by 15 to 20 dB. In contrast, 10-knot downwind propagation (track to site) virtually eliminated all terrain shielding (10-to-15 dB increase). Component of the wind along the track direction appears not important.
Vertical temperature gradients	When aircraft were shielded by terrain, vertical temperature gradients may be associated with increased sound levels of up to 5 dB (during daytime).
Local shielding (such as large boulders)	The association with local shielding appears insignificant, considering all other variability in the measured data.
Effective frequency* for atmospheric absorption	The effective frequency for atmospheric absorption is approximately 350 Hz.

*Effective frequency is derived in Appendix L.7.2.3, page 263.

Insights from these numerical results appear in Section 9.3.4, below.

9.3.3 Sensitivity tests

The best-fit regression equation from this non-linear analysis can also be used to graphically show the apparent effect of several physical factors. In particular, it can be used to test the results of *less-accurate input* on the resulting computations—that is, graphical sensitivity tests. Such use provides insight into the computer models—in particular, when they might be used with less-than-ideal input .

Figure 62 shows the results of less-accurate input. The best-fit comparison appears in this figure's upper-left frame. This plot is identical to Figure 61, except that its regression is linear instead of logistic. The other frames show the result of less-accurate input. Working horizontally from left to right, the frames show the result of less-accurate meteorological input. Working vertically from top to bottom, the frames show the result of less-accurate ambient sound levels and the result of ignoring terrain. Table 34 summarizes the numerical comparisons in this figure.

Table 34. Sensitivity Tests for Less-Accurate Computer Input

Input type	Less-accurate input	Result of less-accurate input (see Figure 62)
Meteorology: temperature and relative humidity	Averaged over all measurements before computation.	No significant result.
Meteorology: vertical temperature gradients and wind	Both omitted.	No significant result.
Ambient sound levels	Tabulated EA ambients instead of ambients measured at specific sites and specific times.	Reduced precision (95-percent confidence ranges are wider).
Terrain	Terrain ignored.	Significant over-computation of audibility.

9.3.4 Model insights from this non-linear analysis

Best-fit values of other regression coefficients can provide insight into discrepancies found above in the computer-model computations. In turn, these insights can suggest model improvements and implications for use of specific input. This section summarizes these insights and implications. Due to the nature of regression analysis, these insights should be thought of as "probable" relationships, not certainties. It should also be noted that because of the complexity of the underlying regression equation, the computer program was not able to estimate the statistical error of the results.

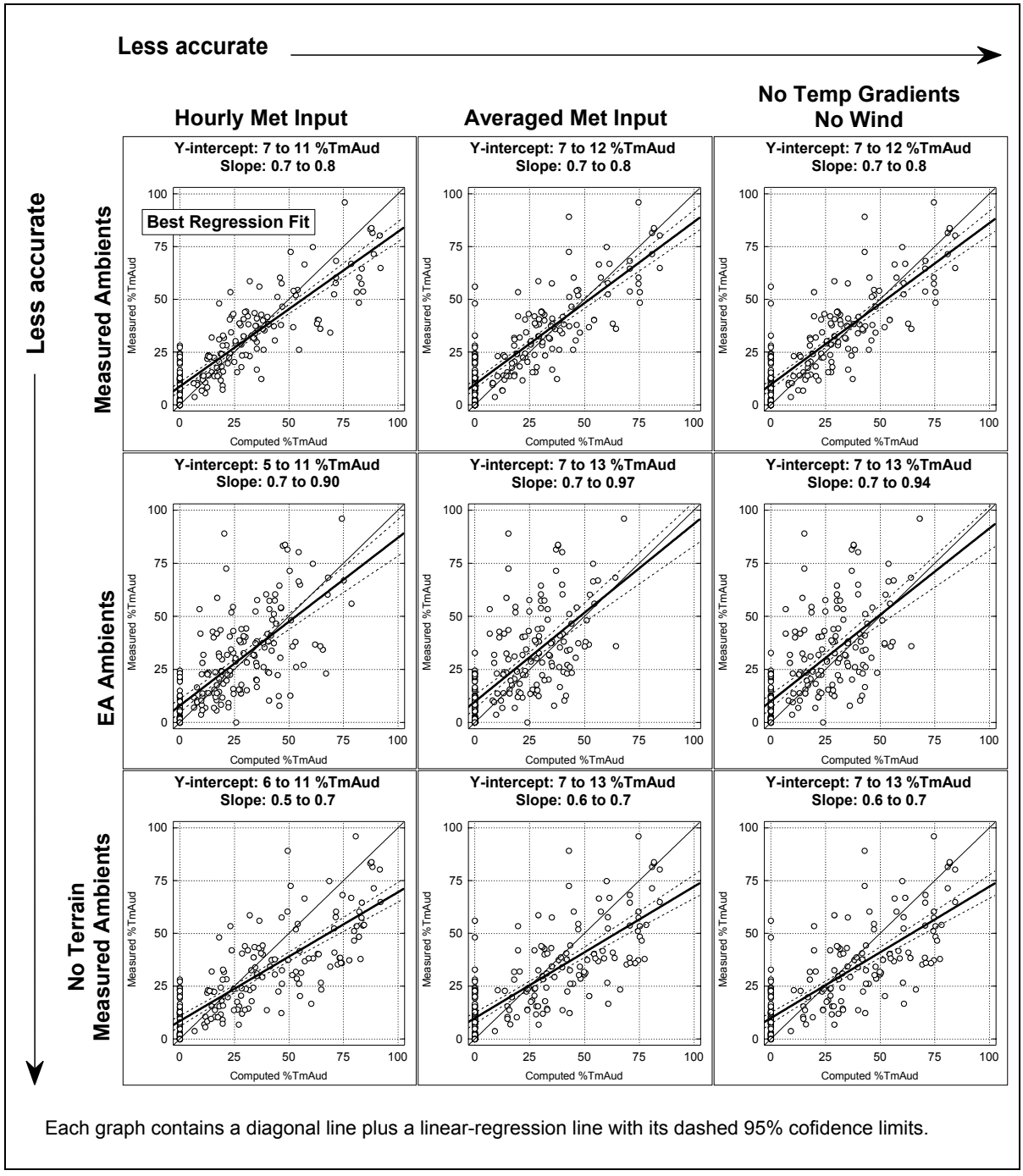


Figure 62. Less Accurate Uses of the Regression Results

From the results in Table 33 and Table 34, the following model insights follow from this analysis:

- *Vistaliners.* Vistaliners produce only 0.3 times the audibility that non-Vistaliners produce, for equal number of each per hour. Therefore, if a tour fleet with *no* Vistaliners were replaced by one of *all* Vistaliners, audibility would be multiplied by 30%—a very large improvement. This conclusion from the non-linear regression can guide decision makers in trying to increase the use of quieter aircraft on air tours above national parks.
- *Terrain shielding.* Terrain shielding greatly affects audibility, by providing an average of 13 decibels of sound-level reduction.[62] This effect does not average out over time. Since the INM versions lack terrain shielding, inclusion of this factor in the model is recommended, Section 11.2.2. Omission of terrain shielding results in significant overcomputation, per the sensitivity tests. See also Section 8.5.3.2 for further discussion.
- *Wind speed and direction, combined with terrain shielding.* Wind can greatly affect audibility from hour to hour. However, these effects would tend to average out over time. None of the computer models includes the propagation effects of wind. Including such algorithms would improve single-hour computations but would probably not improve computation of long-term audibility.
- *Vertical temperature gradients.* Vertical temperature gradients can affect tour audibility from hour to hour, attenuating more in the afternoon when temperature decreases most rapidly with increasing height (greatest temperature "lapse"). None of the computer models includes the propagation effects of vertical temperature gradients. Including such algorithms might improve single-hour computations and might also improve computation of long-term audibility. This conclusion is supported by the sensitivity tests, as well.
- *Local shielding (such as large boulders).* Local shielding is not important to model. Input of detailed shielding due to local boulders appears not necessary.
- *Effective frequency for atmospheric absorption.* The accuracy of INM (A levels) might be improved by computing atmospheric absorption at 350 Hz. This improvement in accuracy does not necessarily mean that sound levels at 350 Hz control audibility. The regression analysis is not powerful enough for such a conclusion.
- *Temperature and relative humidity.* Because the use of the averaged temperature and relative humidity had no significant affect on the relationship of measured to computed values, continued use of averaged input for temperature and relative humidity appears completely adequate.
- *Ambient sound levels.* Use of tabulated ambients (based upon vegetation zones, for example) may decrease the precision of the computations. Most of this increased variability is likely associated with specific sites, rather than specific hours. For this reason, this decreased precision will likely persist in computations of long-term (averaged) audibility.

9.4 Direct Comparison among the Models

In addition to the analysis reported above, the computations from the four computer models were directly compared against each other. Figure 63 through Figure 65show these comparisons.

[62] Non-linear regression involved measured audibility. Nevertheless, the functional form of the regression equation included the effect of intervening terrain as "terrain shielding," in decibels. See Appendix L for equation details.

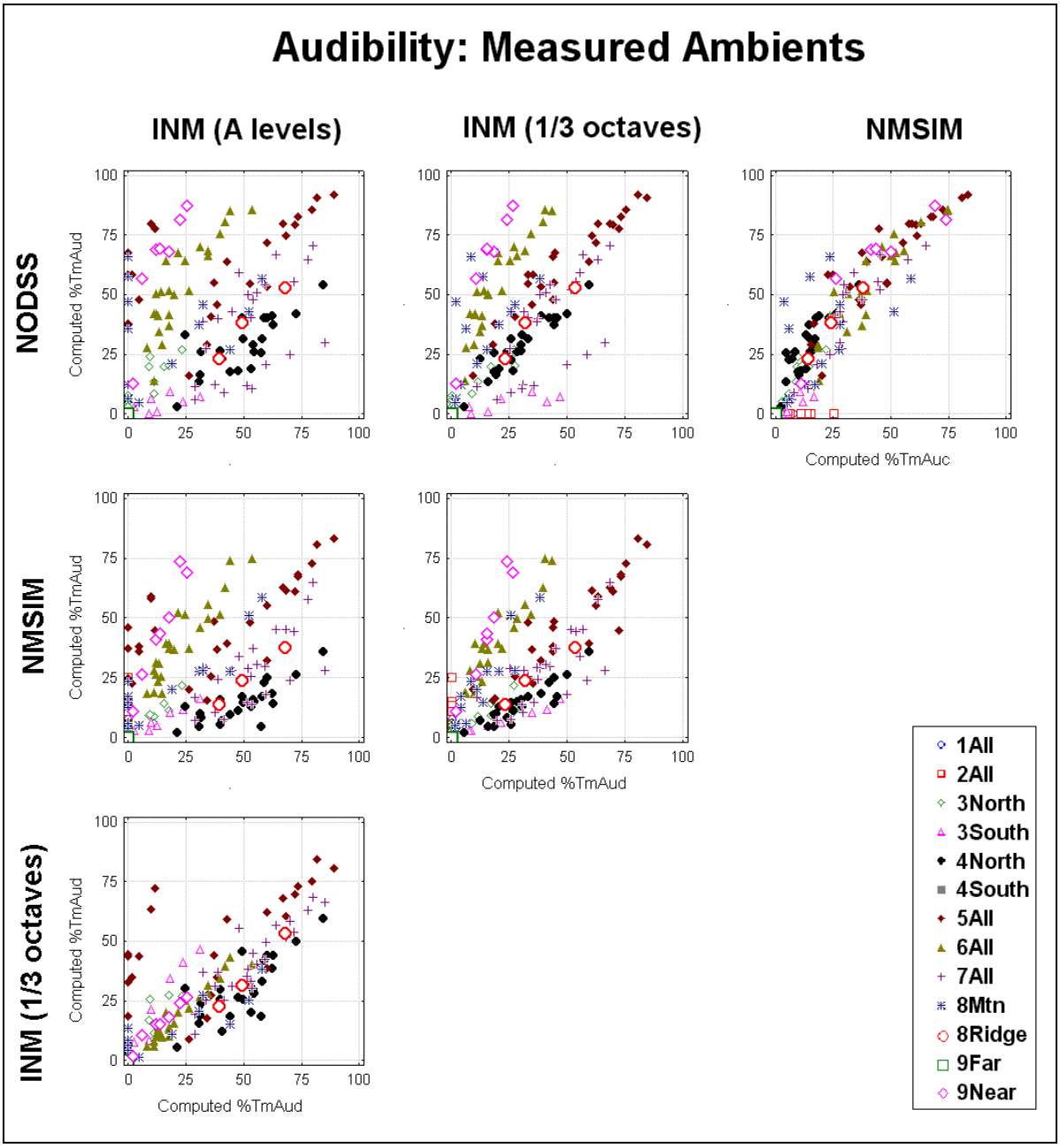


Figure 63. Comparison of Modeled Results: Audibility, Measured Ambients

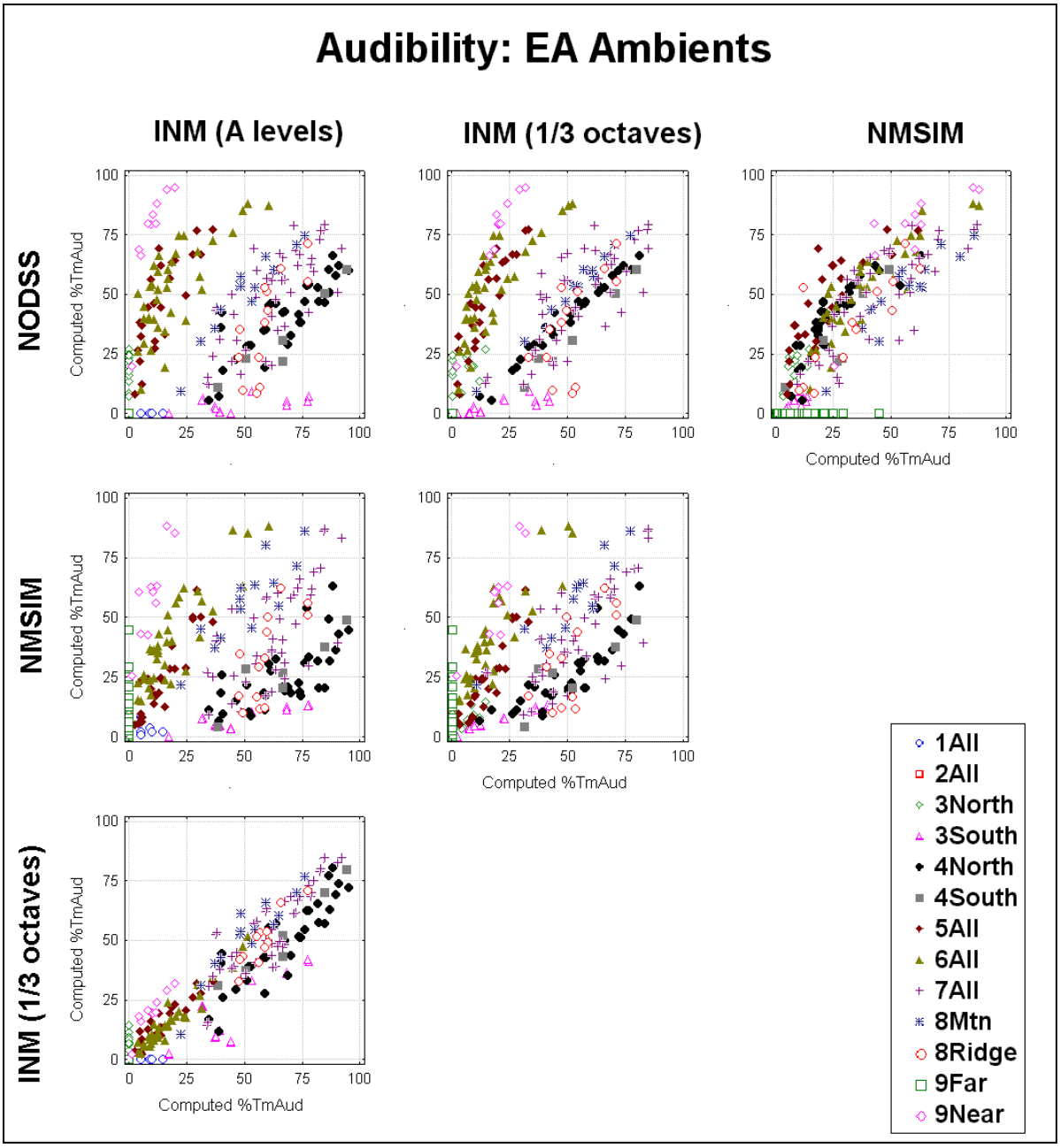


Figure 64. Comparison of Modeled Results: Audibility, EA Ambient

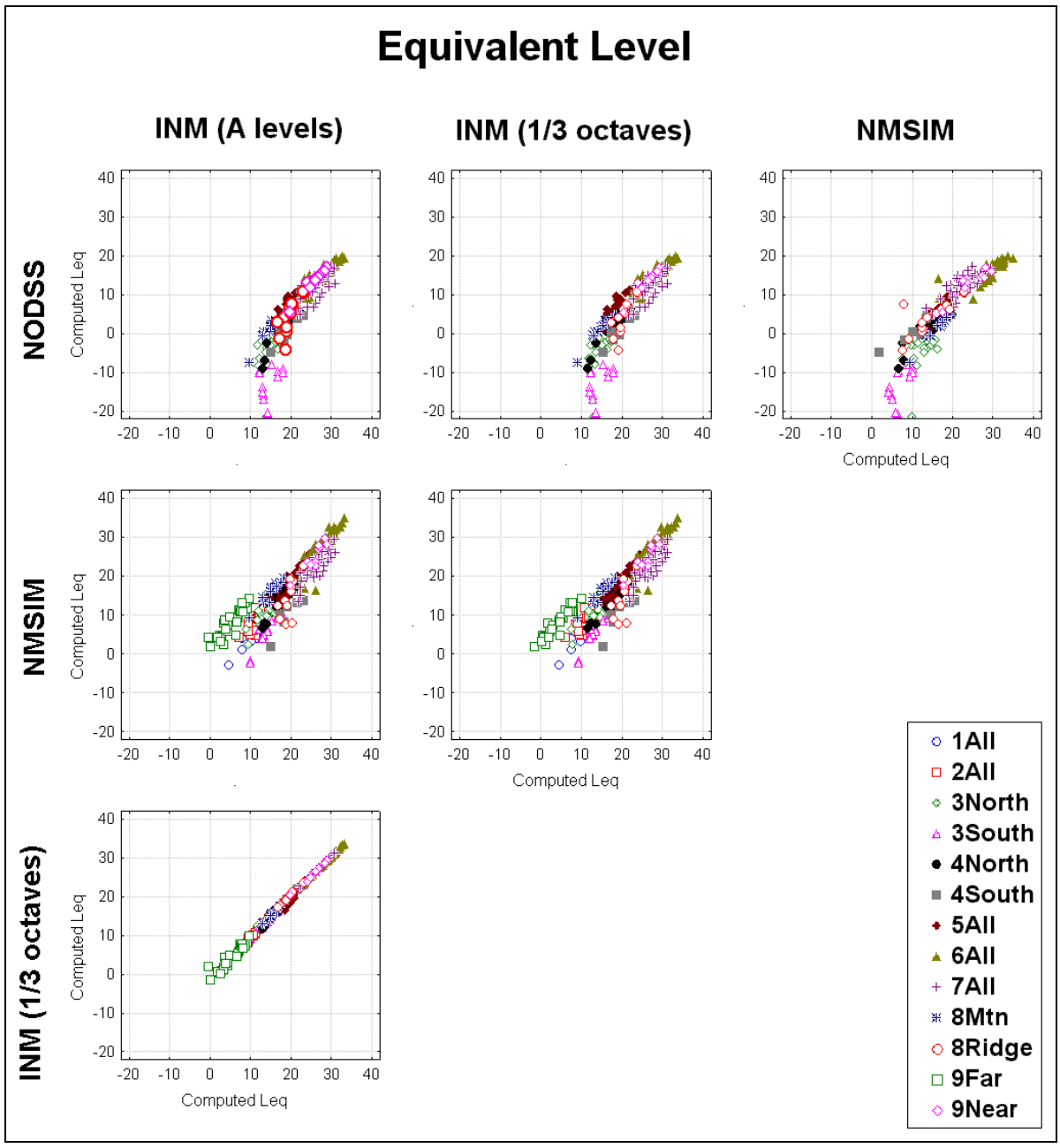


Figure 65. Comparison of Modeled Results: Equivalent Level

These comparisons suggest several observations. First, the computations of the two INM models are similarly affected by the change in ambient levels. The plots of the INM models against each other result in similar distributions. In a like manner, NODSS and NMSIM calculations appear equally affected by the different ambients. Noteworthy, however, are the differences between the two INM versions, on the one hand, and NODSS and NMSIM on the other. Because NMSIM has the smallest error in calculation of audibility, the plots of the INM versions against NMSIM confirm the difficulties the INM versions have when calculating audibility.

Page Intentionally Blank

10. CONCLUSIONS – PREFERRED MODELS

This section presents the conclusions about the models that the authors draw from the analyses presented in this report. It discusses the preferred models for use and the reasons for our preferences. It is a duplicate of Section 1.10.

We consider NMSIM to be the model most suited for use in computing percent of the time tour aircraft are audible. Either version of the INM is suited for computation of hourly equivalent sound levels, and NMSIM performs almost as well. The following paragraphs review the basis of these recommendations.

10.1 Overall Error

For the computation of audibility, NMSIM provides the lowest overall error, whether for the measured or EA ambient, or for the hourly data or the site group data. Additionally, the comparisons of these overall errors for the different ambients and data sets give results that are logical and favorable for use of NMSIM in computations.

The NMSIM overall errors for measured ambients are smaller than for the EA ambients. In these comparisons of measured and computed values, it is useful to keep in mind the differences between the measured and the EA ambients as described in Section 1.9.1.1. The measured ambients were measured at the times when, and at the sites where, the audibility logging was conducted, while the EA ambients are generalized ambients based on earlier data. Hence, use of the EA ambients in computing audibilities should give results similar to those computed using the measured ambients, but with somewhat less accuracy and precision. NMSIM demonstrates this trend.

It is unlikely in future modeling of the Canyon or of other parks that ambient levels will be as widely and thoroughly measured, as were the measured ambients of this study. The ambient levels will have to be generalized from limited measurements.[63] Thus the results using the measured ambients should reveal the "best" that the models can do, given the "best" ambients, while the results using the EA ambients provide what might be considered a more realistic application of the models. The two ambients may be considered as testing the various models' sensitivities to different assumptions about ambient levels, and in this sense can provide additional insight about model performance.

For all models except NODSS, use of measured ambients produces less scatter (less overall error) than use of the EA ambients, and the scatter is in both cases least for NMSIM, and greater for NODSS and for the INM versions.

From this perspective, for audibility, NMSIM provides what we judge to be the best-behaved transition from measured to EA ambient; the data become more scattered, for both hourly and site group data, but still reasonably surround the diagonal of equality. The scatter of the data for the other models changes appreciably from measured ambient to EA ambient, suggesting that the calculations of these other models are more dependent on the specific ambient sound levels that are used.

[63] For example, to model the entire Canyon, generalization of the ambients is necessary and one method is provided in APPENDIX F, page199. It would be valuable to rerun each of the models with these generalized ambients to determine how overall error is affected. Such a run would provide a scenario more typical of an actual park application than that provided by using either the measured or EA ambients, see Section 1.11.3.1.

It is especially desirable that the site group overall error be relatively small. Sites (that is, averages over several hours) are what will generally be used in examining tour operations. First, hour-by-hour operations are unlikely to be known, and in most cases, the goal of modeling will be to examine average operations, rather than the operations of a single specific hour. Second, it is likely that modeling will be used to examine the effects of air tour sounds on specific park locations. Finally, if the model results are to be checked for reasonableness or again validated with measurements, the model with the lowest site error will require the fewest measurement sites. For audibility, NMSIM has the lowest overall site group error.

For computation of hourly Leq, both INM versions have the same and the lowest overall error. Whether for individual hours or for site groups, the INM versions have lower errors than do either NMSIM or NODSS (Table 4). The INM was originally designed primarily for computation of equivalent levels, and the results of this test tend to confirm the versatility of that design for even the complex geometries and terrain of the Canyon.

10.2 Accuracy

Audibility

For the single number bias and confidence ranges (Figure 4 and Figure 6), NMSIM has the narrowest confidence ranges that always include zero (no bias), and a bias that is the same as or smaller than that of the other models (except for EA ambient, Site Groups, where its bias is 2±6 and INM (⅓OB) is 1±13). NMSIM is the model most likely to produce unbiased results. Using the best fit regression line and confidence regions (Figure 12, Figure 13), whether for measured or EA ambient, the NMSIM results agree best with measurements – its regression most closely follows the diagonal, and is closest to it, compared with the other models.

Hourly Equivalent Level

For the single number bias and confidence ranges (Figure 8) the INM versions have the smallest bias with 95% confidence ranges that also includes zero. From the regression fit, both INM versions are equally accurate, and NMSIM slightly less so (Figure 14). NODSS is clearly faulty in its calculations of equivalent levels.

10.3 Precision

In general, precision comparisons among models behave the same as the comparisons of overall error, discussed above in 1.10.1. NMSIM and NODSS have less random error than the INM versions for all percent time audible comparisons, and INM and NMSIM versions have similar random error for hourly equivalent level. For audibility, NMSIM and NODSS have higher correlation coefficients (meaning the model results lie closer to the regression line – have less scatter) than those of the INM versions. For Leq, the INM versions and NMSIM have similar correlation coefficients, while the NODSS coefficient is lower. The corresponding degrees of scatter may be seen in Figure 11 and in Figure 12 through Figure 14.

10.4 Contour error

For many future analyses, one or more of the models will be used to generate contours of equal percent time audible or of equal hourly equivalent level. This analysis estimated the error that is likely to be associated with these contours (see Figure 15 through Figure 17 and Section 8.7, page 119).

Audibility Contours

Since the distance of the contour from the corridor will vary for different corridors, it is desirable for the model's error to be relatively independent of this distance, and as low as possible. NMSIM provides the lowest contour error of the four models, and that error is relatively independent of distance from the corridor.

Hourly Equivalent Contours

Both INM versions and NMSIM compute hourly equivalent level contours with comparable errors. Beyond about 7 miles, the INM error increases to about ± 9 dB to ± 10 dB, while NMSIM error remains at about ± 7 dB, see Section 8.7.5, page126.

10.5 Calibration

Calibration was considered as a possible solution for improving the accuracy of the models. However, not only do we believe that current models are sufficiently accurate for application to parks (however see Section 1.11.2 for areas of possible model improvement), but calibration depends entirely on the available data and makes questionable any wider use of the calibrated model for other park applications.

Page Intentionally Blank

11. RECOMMENDATIONS

This section is a duplicate of Section 1.11.

11.1 Recommended Application of Models

This section presents the authors' recommendations about how the various tested models would be used to achieve the most realistic computed values, based on the results of this study. We realize that both NPS and FAA may have their own requirements and criteria for modeling tour aircraft sounds in parks, and these recommendations are made without consideration of such requirements.

11.1.1 NMSIM

Of the four models, NMSIM is the most likely to compute realistic values of tour aircraft audibility in the Canyon. It can be used to model air tours throughout the entire Canyon by separately modeling twelve to twenty different hours of tour operations randomly chosen from the tour period of interest. The results of these runs should be averaged together, and then audibility contours computed from the averages. Using more than about 12 hours in this process will maximize the probability that the results are realistic, based on the contour error analysis of Section 8.7.4. That section, and Figure 56 and Figure 57 show that the narrowest confidence limits are achieved when many hours of operations are averaged.[64]

NMSIM may be applied to other parks. Though this study has used only Grand Canyon data, the important features of terrain, distance, number of operations, temperature and wind gradients have been included in the analysis and demonstrated no significant biasing of NMSIM results. However, local park ambients should be used, and some type of reasonableness tests of model results should be included for applications to other parks. Ambient levels used will depend upon judgments of what sound levels are appropriate, likely based either upon what ambient sound levels are intruded upon, or on what ambient sound levels affect air tour audibility.[65] These ambients should be adjusted to account for the effects of the human threshold of audibility (see Section 6.1.5.1, page 67). Note that use of NMSIM requires spectral data for both ambient and aircraft sound levels, including directivity information on the aircraft.

Applications to other parks should include tests for "reasonableness" if not strict validation testing. The type of validation provided in this current study is far too demanding of resources to be practical at additional parks. Rather, we propose that 1) careful measurements be made of any tour aircraft used at the park that were not measured in this study and that those measured levels be included in the modeling process; 2) that sound monitoring together with collection of observer logs be done at several sites exposed to tour aircraft noise, and that these measurements be compared with modeled results. Exact procedures for such measurements and comparisons need to be developed.

[64] The data show that with increased number of hours averaged, the 95% confidence limits tend to reduce asymptotically, and above about 12 to 15 hours used for the average, these confidence limits are likely to be within a few percent of the minimum, see for example Figure 58. Naturally, the more hours averaged, the narrower the limits, though with diminishing returns. If the variability in the number of tours per hour during the period of interest is higher than encountered in this study (2 tours per hour to 14 tours per hour), it may be useful to average more hours – perhaps a percent of total hours such as 10%.

[65] This model validation analysis used for the measured ambient, the L_{50}s of periods at each site that the observers identified as natural, see Appendix C.3, page 168. It should be noted that future modeling of the entire Grand Canyon might first be preceded by running the model(s) to be used with the ambient levels derived in APPENDIX F, page 199. This run would show how well the models perform with these new generalized ambients. See also Section 1.11.3.1.

NMSIM may also be used to compute hourly equivalent sound levels for tour aircraft over parks, though the INM versions performed slightly better. Proper spectral data are needed for the aircraft, and reasonableness testing is recommended.

11.1.2 INM, either version

Either version of the INM can be used to compute realistic hourly equivalent sound levels for tour aircraft over the Canyon and for other parks. As discussed in the previous section, 1.11.1.1, several hours of operations (this study suggests more than 10 to 15 hours, see Figure 59) should be randomly selected from the tour period of interest, run in the model, then averaged and used to determine contours, if appropriate. Or equivalently, for hourly L_{eq}, air traffic can be averaged over many hours and then the model run just once. Proper tour aircraft sound level data are needed and, as with NMSIM, reasonableness testing is recommended when the INM is used for other parks.

11.2 Suggested Improvements of Models

Analysis of how physical factors (such as wind speed and direction, ambient levels, etc.) relate to differences between measured and modeled results, as well as analysis of how these factors relate to the measured results, helps to identify which factors may produce model error. Such factors are candidates for inclusion or for further examination in the model. The following suggestions are offered by the authors as initial areas to investigate for improvement and are based on the results of these analyses.

11.2.1 NMSIM

►NMSIM currently does not account for additional attenuation that may result from heavily forested areas. Further development of NMSIM should consider how this additional attenuation, could be included in the model. The analysis showed NMSIM tends to over-predict for these forested areas. This type of attenuation is likely more important for computation of percent time audible than for hourly equivalent levels.

►NMSIM shows a slight bias toward under-prediction of equivalent levels. This under-prediction does not appear to be a result of wind or temperature gradients. Examination of single event sound levels may suggest some possible causes.

►NMSIM generally under-predicts audibility for the "9Near"[66] sites. These sites are about the same distance from the corridor as the 6 and 7 sites, which are not under-predicted. Possibly, the complex flight tracks near the 9Near sites affect NMSIM computations adversely.

11.2.2 INM Models

►Both INM versions compute zero percent time audible for the "9Far"[67] sites when tour aircraft were audible, which suggests these models might be improved through examination of their: 1) assumptions for long-distance propagation, since both models apparently predict levels so low at these distances that they are determined to be inaudible, 2) computation of audibility when aircraft

[66] 9Near sites are sites 9C and 9F, Table 20, page 88, which are about 2 miles from the corridor (see Figure 22, page 54 and Table 11, page 55).
[67] 9FAR sites are sites 9A, 9B, 9D and 9E, Table 20, which were 11 to 15 miles from the flight corridor (see Figure 22, page 54 and Table 11, page 55).

sound levels are low, and 3) computation when only a small portion of a flight track contributes to the sound levels.[68]

►Both INM versions uniformly underestimate time audible at 9Near sites, suggesting that how these models treat curved flight tracks might be examined, since these sites are likely to receive sound from several portions of the track that curves into and out of the Little Colorado.

►The INM over-predicts audibility close to the corridor (within 0 – 6 miles) when shielding is present (visible angle is small), but under-predicts at these distances when little shielding is present (visible angle is large), see Figure 44, page 105 and Figure 45, 106. The former result is likely due to the fact that the INM does not account for the shielding effects of terrain, while the latter effect may be the result of how the model treats the various parameters associated with audibility, such as the source directivity assumptions. Hence, inclusion of terrain shielding should be considered. Also, for the INM 1/3 octave band version, the components of audibility calculations, especially source directivity should be examined.

►As with NMSIM, the INM versions do not include attenuation of tour aircraft sound levels due to expanses of forested areas. The analysis showed that the INM ⅓ octave band model tends to over-predict audibility for these areas.

11.2.3 NODSS

►NODSS computations of equivalent levels should be examined. All NODSS results show a clear bias toward under-prediction of hourly equivalent results. NODSS was designed to compute *total* hourly equivalent level, including the contribution of the natural ambient. Since such results would not provide an appropriate comparison with measured results, NODSS input was modified, see Section 3.4.2. This modification may have caused the significant under-prediction of computed equivalent levels, though currently, no explanation has been determined.

►NODSS also appears to over-predict audibility in the forested areas. Inclusion of adding this type of attenuation should be considered.

►NODSS computes zero audibility for the distant 9Far sites where aircraft were audible. As with the INM versions, reasons for this under-prediction should be examined.

11.2.4 Factor Not Recommended for Inclusion

One factor that may have some significance, vertical temperature gradient, is not recommended for inclusion in any of the models. Though absence of this factor in the models could result in some over-prediction at large distances, particularly with respect to equivalent levels, see Figure 47 through Figure 49, the complex relationships between this factor, distance and terrain shielding makes derivation of the exact importance of this factor virtually impossible with the current data. Moreover, in terms of audibility, all models tend towards slight under prediction at these larger distances, so that the net effect of temperature gradient as evidenced by the available data suggests that temperature gradient did not have a dominant effect on the measured results. Finally, acquisition of temperature gradient information for incorporation in future modeling, whether at the Canyon or other parks, is likely to be well beyond the resources available for data collection.

[68] From the location of 9Far sites, the flight corridor would subtend a relatively small angle, less than 45 degrees.

11.3 Suggested Possible Further Analysis

11.3.1 Run Models Using Generalized Ambients

New generalized ambient levels have been developed that can be used throughout the Canyon.[69] APPENDIX F, page 199, provides the derivation of these ambients. At the discretion of NPS / FAA, these values may be used first to run any of the models that will be used to compute audibility Canyon wide. Results would be compared with measured audibilities, and model performance determined. Such application will provide a realistic assessment of how well the models perform when carefully measured, but generalized ambients are used.[70] This approach recognizes that ambients similar to the "measured ambients" used in this study will rarely, if ever, be available for modeling purposes. After this run and analysis, model performance using the generalized ambients will be known.

11.3.2 Additional Analysis of Quiet Aircraft

One of the primary reasons for conducting the regression analysis of the measured results was to determine whether quieter aircraft, such as the Vistaliner (a specially quieted Twin Otter / DHC6 using Raisbeck designed modifications to the fuselage and quiet propellers) could have a statistically measurable effect on tour aircraft audibility (see Section 9.3). The analysis shows that the Vistaliner audibility was, on average, 30% that of other tour aircraft. (See Table 17, page 70 for a complete list of aircraft types measured.) If aircraft like the Vistaliner replaced the other aircraft measured here, they would very significantly reduce audibility of tour aircraft in the Canyon.

FAA has a congressional mandate to identify "quiet technology" aircraft that could be used as tour aircraft. Congress has designated such quiet aircraft as eligible for special consideration in use of tour routes over national parks.[71] It is possible that the resulting FAA research efforts on quiet technology aircraft could benefit from further detailed analysis of this study's data to determine whether the tour aircraft types might be rank-ordered by their relative contributions to audibility. Such rankings might be useful in FAA's efforts to define "quiet technology" as required by law.

11.3.3 Model Testing Procedure

The National Parks Air Tour Management Act of 2000 establishes a public process for development of Air Tour Management Plans (ATMP's). It is likely that the ATMP development process will require modeling of tour aircraft at other National Parks. For these applications, it will be useful if some basic procedures are defined for testing the reasonableness of the modeled results for the park under examination. Procedures would include methods for measurement of aircraft types not already measured for the present study, and collection of data for comparison with model results.

11.3.4 Computation of Parkwide Metric Error

It is likely that any single-number, parkwide impact metric will have considerably lower overall error than the values reported here for hourly or site error. A parkwide metric is an average over a large

[69] These generalized ambients are similar in concept to the EA ambients; they apply throughout the Canyon based on vegetation zone. These new generalized ambients, however, are derived from the data acquired as part of this study, unlike the EA ambients that were derived from previous measurements, see references in footnote 43 page 59.

[70] The authors judge, however, that the performance would likely be between that found in this study using the measured ambients and that found using the EA ambient.

[71] Title VIII of Public Law 106-181, National Parks Air Tour Management Act of 2000.

number of computed sites throughout a park. Hence, it can average out both the hour-to-hour error, and the site-to-site error reported here. Over-predicted sites tend to balance out under-predicted ones.

An important single-number parkwide metric in the Canyon is the computed fraction of land area where tour aircraft are audible more than 25% of the time. This study's computer programs could be used to compute this value, and this study's results then used to determine the confidence interval of this single-number metric of impact. We suggest that this computation of error be done, due to the importance of this metric in determining restoration of natural quiet in the Canyon. By applying propagation of error techniques (mathematical methods that combine the uncertainties of multiple factors into the resulting uncertainty of a single function of those factors) to the site errors for each model, it would be possible to estimate the error associated with a model-computed area exposed to tour audibility in more than 25% of the time.

11.3.5 Use Measured Data to Test Detection Algorithms

The measured data (which includes second-by-second 1/3 octave band levels and associated second-by-second observer logs) represents virtually the best data source possible for testing automated identification of "natural" and "aircraft" sound levels. Ultimately, most sound measurements in parks will probably need to be collected with unattended, long-term monitoring. It will be extremely advantageous if these unattended data can be reliably used to quickly determine the sound levels of the natural ambient and the number and sound levels of intrusions. The measured data collected for this study provide the means for testing and checking the reliability of various detection algorithms with respect to human determination of audibility.

11.3.6 Rerun NMSIM with Equally or Randomly Spaced Aircraft

For the study, NMSIM ran the aircraft flights with the actual timings that they flew. In modeling of future studies at other parks, the exact timing and spacing of tours will probably not be known. The model could be run with aircraft at equal spacings and at random spacings to determine the magnitude of the error such approximations can produce. These runs could also help determine how best to select tour aircraft spacings for modeling when the actual spacings are unknown.

11.3.7 Revise "Compression" Algorithm

Neither the INM versions nor NODSS account for the overlapping of aircraft audibility when aircraft fly in close succession. These models compute the audibility duration for each aircraft separately, and then add all durations together. Such an approach will over-predict total audibility when aircraft fly close enough to result in audibility of more than one aircraft at a time. A "compression" algorithm was derived empirically from previous measurements to reasonably reduce these computed audibilities, see APPENDIX J, page 243. The data from this current study could be used to develop an up-dated compression algorithm that might be applicable to more situations and more parks.

Page Intentionally Blank